Available online at www.sciencedirect.com

SCIENCE DIRECT®

Discrete Applied Mathematics 154 (2006) 1100–1112

DISCRETE APPLIED MATHEMATICS

www.elsevier.com/locate/dam

Subset-conjunctive rules for breast cancer diagnosis

Rajeev Kohli[a], Ramesh Krishnamurti[b,1], Kamel Jedidi[a]

[a] *Graduate School of Business, Columbia University, USA*
[b] *School of Computing Science, Faculty of Applied Sciences, Simon Fraser University, 8888 University Drive, Burnaby, B.C., Canada V5A 1S6*

Received 12 May 2003; received in revised form 5 August 2004; accepted 19 October 2004
Available online 28 November 2005

Abstract

The objective of this study was to distinguish within a population of patients with and without breast cancer. The study was based on the University of Wisconsin's dataset of 569 patients, of whom 212 were subsequently found to have breast cancer. A subset-conjunctive model, which is related to Logical Analysis of Data, is described to distinguish between the two groups of patients based on the results of a non-invasive procedure called Fine Needle Aspiration, which is often used by physicians before deciding on the need for a biopsy. We formulate the problem of inferring subset-conjunctive rules as a 0–1 integer program, show that it is NP-Hard, and prove that it admits no polynomial-time constant-ratio approximation algorithm. We examine the performance of a randomized algorithm, and of randomization using LP rounding. In both cases, the expected performance ratio is arbitrarily bad. We use a deterministic greedy algorithm to identify a Pareto-efficient set of subset-conjunctive rules; describe how the rules change with a re-weighting of the type-I and type-II errors; how the best rule changes with the subset size; and how much of a tradeoff is required between the two types of error as one selects a more stringent or more lax classification rule. An important aspect of the analysis is that we find a sequence of closely related efficient rules, which can be readily used in a clinical setting because they are simple and have the same structure as the rules currently used in clinical diagnosis.

Keywords: Classification; Data mining; Subset conjunction; Logical Analysis of Data; Partially defined Boolean functions; Integer programming; Approximation algorithms; Computational complexity; Cancer; Diagnostic rules; Risk prediction

1. Introduction

In recent years, there has been a dramatic increase in the use of technology in medicine. The complexity and sophistication of the technologies often requires the solution of decision problems using combinatorics and optimization methods [15]. Logical Analysis of Data (LAD) is one such method; it infers logical classification rules from outcome data [1,6,7,11]. The purpose of this paper is to describe a subclass of LAD models related to the types of rules

E-mail address: ramesh@cs.sfu.ca (R. Krishnamurti).

[1] Partly supported by a research grant from the Natural Sciences and Engineering Research Council of Canada.

0166-218X/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.dam.2004.10.010

clinicians use for medical diagnosis. We formulate the model as a 0–1 integer program, describe its relation with the satisfiability problem, show that it is NP-Hard, examine the theoretical properties of deterministic and probabilistic approximation algorithms, and describe an application to the diagnosis of breast cancer using the Wisconsin breast-cancer data.

We call the proposed models *subset-conjunctive models.* They are generalizations of conjunctive and disjunctive decision models in psychology (see, e.g., [8]). These rules are known to be often used by people in phased-decision making (e.g., [12,16,19–22,26,28,31]). Examples of the use of these rules are the formation choice or consideration set of brands by consumers [2,13,23,24]; the specification of target markets by marketing managers; and the screening of applicants for jobs, loans and school admissions (e.g., [10,12]).

The proposed model is also related to Boolean regression [4,5] and to methods for identifying partially defined Boolean functions [9]. As Crama et al. [9] note, there can be a large number of possible extensions for constructing the set of all partially defined Boolean functions. To reduce the problem to a manageable size, it becomes useful to restrict the analysis to a subclass of functions. In the present model, we investigate a subclass of functions, which does not necessarily classify every outcome correctly, for two reasons. One is the presence of error in data, which can arise, for example, when the symptoms are indicative of a disease but not always sufficient to make a certain diagnosis. The other is the implicit hypothesis that the underlying function classifies outcomes based on the number of symptoms indicative of disease: if the presence of k symptoms in a patient is indicative of a disease, then so is the presence of $k+1$ such symptoms. The set of symptoms indicating disease corresponds to the set of literals set true; the objective is to identify disease indicators and a subset size k to maximize the number of patients correctly classified. The problem we address here is itself more general: each attribute can have more than two levels (thus each variable is in general n-ary, as opposed to binary), and differential weights can be assigned to misclassifications of patients with and without a disease.

The subset-conjunctive rules we describe can be useful in situations where any one single piece of information is not conclusive for making a decision or classifying an outcome, and it is also not possible to obtain all relevant information. Such situations often arise in medical diagnosis when a patient has some but not all possible symptoms of a disease, either because not all symptoms may ever manifest themselves or because some symptoms only appear in later stages of affliction. Second, subset-conjunctive rules offer flexibility to a decision maker, in the sense that these can be fine-tuned to fit the risk attitudes of a physician. To illustrate, consider the screening of breast-cancer patients for a biopsy, a clinical procedure that is both traumatic and expensive. Most patients considered for a biopsy show several symptoms of breast cancer. But restricting a biopsy to only those patients who have all possible symptoms may be too conservative (in the sense that only those who are virtually certain to have breast cancer will get a biopsy); and recommending a biopsy to patients who have any one symptom may be too permissive (in the sense that almost everyone will get a biopsy). A subset-conjunctive rule can be useful because it allows one to adjust the number of criteria for screening patients for a biopsy. As one changes the size of the subset over which a conjunction is required, one gets a tradeoff between the numbers of false negatives and false positives. An enumeration of these rules can then help a clinician explicitly assess the tradeoff in the two types of errors across the rules; and can help decide which additional symptom to consider or relax when making the rule more or less lax. A nice feature of these rules is that they have precisely the same form as current diagnostic rules used by physicians: a patient is diagnosed to be at risk if he/she has at least a minimum number of disease symptoms. The difference is that here the rules are based on a model that searches among all possible rules, selecting those that are the most efficient in the sense that they form a Pareto optimal set of rules that tradeoff type I and type II errors of diagnosis.

We formulate the problem of inferring subset-conjunctive rules as a 0–1 integer program, show that it is NP-Hard, and prove that it admits no polynomial-time constant-ratio approximation algorithm. We examine the performance of a randomized algorithm in which each variable is set to be an indicator of a disease with probability 1/2; we also examine the performance of a randomized algorithm using LP rounding. In both cases, the expected performance ratio is arbitrarily bad. We then examine a problem in breast-cancer diagnosis using the Wisconsin breast-cancer data [17]. We use a deterministic greedy algorithm to identify a Pareto-efficient set of subset-conjunctive rules. We describe how the rules change with a re-weighting of the type-I and type-II errors; how the best rule changes with the subset size; and how much of a tradeoff between the two types of error is required by selecting a more stringent or more lax classification rule. A key aspect of the analysis is that we find a sequence of closely related efficient rules, which can be readily used in a clinical setting because they are so simple and have the same structure as the rules currently used in clinical diagnosis.

2. Model formulation

Let m denote the number of criteria or attributes used to make a particular diagnosis. Let attribute j have n_j possible values or levels, $1 \leqslant j \leqslant m$. For example, cell size is an attribute in our application, and "large," "medium" and "small" are its levels. We only consider discrete attributes with a finite number of levels. Each attribute level is either indicative of a disease, or it is not. We will call levels that indicate a disease "acceptable" levels and those that do not indicate a disease "unacceptable" levels. Each attribute level is assumed to be acceptable or unacceptable independently of other attribute levels.

A "profile" refers to the description of a patient in terms of a subset of attributes; each profile has at least one and at most m attribute levels. Let C denote the set of profiles. Each profile belongs to one of two subsets $A \subset C$ or $U = C \backslash A$. For example in our application, A and U are the sets of profiles for patients with and without breast cancer. For succinctness, we call A the set of acceptable profiles and U the set of unacceptable profiles.

A subset-conjunctive rule classifies a profile as being acceptable if it has at least $1 \leqslant k \leqslant m$ acceptable attribute levels; otherwise, it classifies the alternative as unacceptable. Evidently, $k = 1(m)$ correspond to disjunctive (conjunctive) classification of the profiles. If all levels of an attribute are acceptable, then the attribute is irrelevant to the evaluation and a subset-conjunctive rule requiring the satisfaction of k criteria is in fact equivalent to a rule requiring the satisfaction of $k-1$ criteria for the diagnosis of a disease. We therefore require that at least one level of each attribute is unacceptable.

Let $y_i = 1$ if profile $i \in A$ is correctly classified (as having the disease) by a subset-conjunctive rule; otherwise, let $y_i = 0$. Similarly, let $z_i = 1$ if profile $i \in U$ is incorrectly classified (as having the disease) by a subset-conjunctive rule; otherwise, let $z_i = 0$. Let $x_{ijp} = 1$ if level p of attribute j appears in profile i, $1 \leqslant p \leqslant n_j, 1 \leqslant j \leqslant m, i \in C$. Let $\beta_{jp} = 1$ if level p of attribute j is predictive of a disease; otherwise, $\beta_{jp} = 0$. Suppose we set the subset size k to a pre-selected value $1 \leqslant k \leqslant m$. That is, we tentatively specify the minimum number of predictive symptoms a patient must have before being classified as having a disease. The solution to the following integer program identifies which attribute levels are predictors of a disease and which are not; i.e., it solves for a vector $\beta = \{\beta_{jp} | 1 \leqslant p \leqslant n_j,\ 1 \leqslant j \leqslant m\}$, that maximizes the difference between the weighted number of correct classifications of patients $i \in A$ (classifying as those who have the disease) and the weighted number of incorrect classifications (as those who have the disease) among the patients $i \in U$ who do not have the disease. The weights w_A and w_U reflect the relative importance of correctly (incorrectly) classifying patients in group A (group U). Thus, $w_U = 0$ $(w_A = 0)$ means that a physician is only concerned with identifying patients with (without) a disease, but is not concerned if patients without (with) the disease are incorrectly classified. Typically, it makes sense to have $w_A > w_U$, where A comprises the set of individuals who have a disease. The choice of w_A and w_U is, in our opinion, best made by a physician; our model should facilitate the decision by providing alternative rules for different values of these weights.

Problem P.

$$\text{Maximize:} \quad \mathbf{Z}_k = w_A \sum_{i \in A} y_i - w_U \sum_{i \in U} z_i$$

subject to:

$$\sum_{j=1}^{m} \sum_{p=1}^{n_j} \beta_{jp} x_{ijp} \geqslant k y_i \quad \text{for all } i \in A,$$

$$\sum_{j=1}^{m} \sum_{p=1}^{n_j} \beta_{jp} x_{ijp} \leqslant m z_i + (k-1) \quad \text{for all } i \in U,$$

$$\beta_{jp} = 0, 1 \text{ (integral)} \quad \text{for all } 1 \leqslant p \leqslant n_j,\ \ 1 \leqslant j \leqslant m,$$

$$y_i = 0, 1 \text{ (integral)} \quad \text{for all } i \in A,$$

$$z_i = 0, 1 \text{ (integral)} \quad \text{for all } i \in U.$$

The first constraint requires that patient (profile) $i \in A$ be classified as having a disease if at least k of the criteria are indicative of the disease. The second constraint requires that patient (profile) $i \in U$ be misclassified as having the disease if at least k of the criteria are indicative of the disease. The third constraint is an integrality restriction that assures that $\beta_{jp} = 1(0)$ if level p of attribute j is indicative (not indicative) of a disease, $1 \leqslant p \leqslant n_j, 1 \leqslant j \leqslant m$. The fourth (fifth) constraint ensures $y_i = 1$ $(z_i = 1)$ if profile $i \in A$ $(i \in U)$ is correctly classified (incorrectly classified) by an assignment β. The optimal value of k, the subset size, is obtained by solving the problem for different values of $1 \leqslant k \leqslant m$, and selecting the value for which the largest value of Z_k is obtained. If the levels of an attribute are a priori ordered—for example, if a larger lump is no less indicative of breast cancer than a smaller lump—then we can add additional constraints of the type $\beta_{jp_1} \geqslant \beta_{jp_2}$, when level p_1 is a priori no less likely to be indicative of a disease as level p_2, both of which are possible values for attribute j, $1 \leqslant j \leqslant m$.

3. Complexity

The above optimization problem is NP-Hard. This may be seen by considering the case $n_j = 2$, for all $1 \leqslant j \leqslant m$. For each attribute, we assume that one level is indicative of a disease and the other is not; otherwise, an attribute is not useful for discriminating between the two classes of patients, and we can remove it from further consideration as a classification variable. For the remaining attributes, the problem of identifying attribute levels that are indicative of a disease corresponds to a variant of the maximum satisfiability (maxsat) problem in which a clause $i \in C$ is satisfied if it has at least k true literals, $k \geqslant 1$. The problem is to find a truth assignment that maximizes the number of clauses satisfied in a subset $A \subseteq C$ minus the number of clauses satisfied in $U = C \backslash A$. We call this the *maximum subset-conjunction problem*; it reduces to the maxsat problem when $k = 1$ and $C = A$, and to the minsat problem [14] when $k = 1$ and $C = U$. Consequently, maximum subset-conjunction is NP-Hard. From the discussion below, it also follows that unless P = NP, all deterministic approximation algorithms for the problem perform arbitrarily badly in the worst case.

Consider $A = C$ in the maximum subset-conjunction problem; we call this special case *conjunctive maxsat* because a clause is satisfied only if each literal it contains is satisfied. Any instance of conjunctive maxsat where each clause contains $k_i \leqslant l$ literals can be trivially transformed to an instance of conjunctive maxsat where each clause contains l literals (this can be done by augmenting clause i with $l - k_i \geqslant 0$ new literals). Such an instance is equivalent to a posiform representation of a pseudo-boolean function, each clause being associated with a term in the posiform. For a discussion of posiform representations of pseudo-boolean functions, see [3]. Obtaining a truth assignment that maximizes the number of clauses satisfied in an instance of conjunctive maxsat is then equivalent to obtaining a binary vector that maximizes the corresponding posiform function. Conjunctive maxsat is thus equivalent to posiform maximization, which in turn is equivalent to determining the maximum independent set of a graph [3]. It follows that like the latter problem, maximum subset-conjunction has no constant-ratio approximation algorithm unless P = NP.

4. Algorithms

We examine a class of greedy solution procedures for maximum subset-conjunction in this section. The general form of the algorithm is as follows. We start with a feasible, possibly random assignment vector

$$\beta^0 = \{\beta_1^0, \beta_2^0, \beta_3^0, \ldots, \beta_j^0, \ldots, \beta_m^0\}.$$

Let β^0 satisfy G_A^0 clauses in A and G_U^0 clauses in U. Let $G_T^0 = G_A^0 - G_U^0$. We generate a solution vector $\beta^{0'}$ that modifies β^0 by replacing β_j^0 by $1 - \beta_j^0$. Let $G_T^{0'} = G_A^{0'} - G_U^{0'}$ denote the difference between the number of clauses satisfied in sets A and U by the assignment vector $\beta^{0'}$. Let

$$\beta^{(1)} = \begin{cases} 1 - \beta^0 & \text{with probability } p(G_T^{0'}, G_T^0), \\ \beta^0 & \text{otherwise,} \end{cases}$$

where $0 \leqslant p(G_T^{0'}, G_T^0) \leqslant 1$ is a probability that is a function of $G_T^{0'}$ and G_T^0.

In general, let

$$\beta^s = \{\beta_1^s, \beta_2^s, \beta_3^s, \ldots, \beta_j^s, \ldots, \beta_m^s\}, \quad 0 \leqslant s \leqslant m - 1,$$

denote the vector of truth assignments at step s of the algorithm. Let

$$\beta^{(s+1)} = \begin{cases} \beta^{s'} & \text{with probability } p(G_T^{s'}, G_T^s), \\ \beta^s & \text{otherwise,} \end{cases}$$

where $\beta^{s'}$ denotes a solution vector that modifies β^s by replacing β^s_j by $1-\beta^s_j$; G^ℓ_A (G^ℓ_U) denotes the number of clauses that assignment ℓ satisfies among the clauses in A (U); $G^\ell_T = G^\ell_A - G^\ell_U$, $\ell = s, s'$; and $0 \leqslant p(G^{s'}_T, G^s_T) \leqslant 1$ is a probability that is a function of $G^{s'}_T$ and G^s_T.

Different functional forms for the probability give different rules. A deterministic greedy heuristic corresponds to

$$p(G^{s'}_T, G^s_T) = \begin{cases} 1 & \text{if } G^{s'}_T \geqslant G^s_T, \\ 0 & \text{otherwise.} \end{cases}$$

One probabilistic algorithm is obtained by setting

$$p(G^{s'}_T, G^s_T) = \frac{G^{s'}_T}{G^{s'}_T + G^s_T}.$$

This algorithm replaces β^s_j by $1-\beta^s_j$ with a probability proportional to the value of $G^{s'}_T$. Another probabilistic rule, which gives disproportionately greater weight to the better solution is

$$p(G^{s'}_T, G^s_T) = \frac{e^{kG^{s'}_T}}{e^{kG^{s'}_T} + e^{kG^s_T}} = \frac{1}{1 + e^{k(G^s_T - G^{s'}_T)}}, \quad k \geqslant 0.$$

The special case $k=0$ gives $p(G^{s'}_T, G^s_T) = 1/2$, a random algorithm. As k becomes larger, the value of $p(G^{s'}_T, G^s_T)$ goes to zero or one and the algorithm approaches a deterministic greedy heuristic in the limit. We say that an algorithm completes a single "pass" after m steps; i.e., after each β variable has been examined once. In practice, one should run the algorithm for a large number of passes. It is also appropriate to implement the algorithm using several different starting solutions.

Note that the result of the last section implies that unless P = NP, there can be no deterministic algorithm with a constant performance ratio for maximum subset-conjunction. For example, consider the use of the deterministic greedy heuristic on a problem instance with $A = \phi$, $U = C$, with $|C| = |U| = m+1$ and subset size $k=1$ (disjunction). Clause 1 is $\beta_1 \vee \beta_2 \vee \cdots \vee \beta_m$; clause i, $2 \leqslant i \leqslant m+1$, is $\bar{\beta}_{i-1}$. If the algorithm starts with the assignment $\beta_1 = \beta_2 = \cdots = \beta_m = 0$, then there is no improvement in the first pass, and the algorithm returns the same assignment, satisfying m clauses. The optimal assignment is given by $\beta_1 = \beta_2 = \cdots = \beta_m = 1$ and the optimal number of clauses satisfied is 1. Thus, the heuristic has an arbitrarily bad worst-case performance ratio.

Although deterministic algorithms must be arbitrarily bad, it is possible that there are randomized algorithms for the problem that have an expected performance ratio that is not arbitrarily bad. We show below that the expected performance ratio is arbitrarily bad for both random assignments ($p(G^{s'}_T, G^s_T) = 1/2$) and randomized rounding.

4.1. Random assignment

The expected number of clauses satisfied by a random assignment for the maxsat problem is at least 1/2 the total number of clauses [18]. We consider the expected number of clauses satisfied by a random assignment of truth values to variables. Let the number of literals in clause i be $k_i \geqslant k$, where it is required to set at least k variables to true for the clause to be satisfied. For the random assignment where each variable is set to true with probability 1/2 (and false with probability 1/2), the probability of satisfying the clause is given by $1 - [2^{-k_i} \sum_{j=0}^{k-1} \binom{k_i}{j}]$. The worst case occurs when $k = k_i$, when the probability of satisfying clause i drops to 2^{-k_i}. Noting that the expected number of clauses satisfied by a random assignment equals the sum of the expectation of each clause being satisfied, the expected performance ratio of the random assignment becomes arbitrarily bad as k_i goes to infinity.

4.2. Randomized LP rounding

Another randomized algorithm for maxsat is LP rounding [27]. We consider LP rounding for maximum subset-conjunction. We relax the integrality condition on the variables in problem **P** and solve it using linear programming. We let $\hat{\beta}_j$, $\hat{z}_i$ denote the optimal solution to the linear program for the variables β_j, z_i, $1 \leqslant j \leqslant m$, $i \in C$. We then set literal b_j to be true with probability $\hat{\beta}_j$, $1 \leqslant j \leqslant m$. How well will this procedure do on average? We consider the special cases $U = \emptyset$ and $A = \emptyset$; in each case, the lower bound on the expected performance ratio is arbitrarily bad.

We consider $U = \emptyset$; i.e., a version of maximum satisfiability in which a clause is satisfied if it has at least k true literals. A similar analysis can be used to show that the expected performance ratio of the maximum subset-conjunction is arbitrarily bad when $A = \phi$ and $U = C$.

Without loss of generality, consider a clause c_i with all positive literals. Then

$$p(c_i \text{ satisfied}) = \prod_{j \in C_i^+} \hat{\beta}_j.$$

The minimum value of this probability, subject to the constraint

$$\sum_{j \in C_j^+} \hat{\beta}_j \geqslant k\hat{z}_i,$$

is

$$p(c_j \text{ satisfied}) = \begin{cases} 0 & \text{if } \hat{z}_i \leqslant (k-1)/k, \\ k(\hat{z}_i - 1) + 1 & \text{if } \hat{z}_i > (k-1)/k. \end{cases}$$

This may be seen by noting that when $\hat{z}_i \leqslant (k-1)/k$, then $\prod_{j=1}^{k} \hat{\beta}_j$ is minimized by letting at least one of $\hat{\beta}_j$'s be 0. When $\hat{z}_i > (k-1)/k$, then $\prod_{j=1}^{k} \hat{\beta}_j$ is minimized by letting $k-1$ of the $\hat{\beta}_j$'s be 1, and the remaining $\hat{\beta}_j$ take on the value $k\hat{z}_i - (k-1)$. It follows that the expected performance ratio is 0 when $\hat{z}_i \leqslant (k-1)/k$. However, when $\hat{z}_i > (k-1)/k$, then the performance ratio is given by

$$\frac{\sum_{i=1}^{n} k(\hat{z}_i - (k-1)/k)}{\sum_{i=1}^{n} \hat{z}_i} = k - \frac{n(k-1)}{\sum_{i=1}^{n} \hat{z}_i}.$$

For $\hat{z}_i > (k-1)/k$, $i \in C$, the expected performance ratio goes up linearly to 1 as $\hat{z}_i$, $i \in C$, go to 1. Thus, on average, the expected performance ratio of the randomized algorithm using LP rounding is arbitrarily bad.

Given that none of these methods is superior from a theoretical standpoint, we restrict ourselves to the deterministic greedy algorithm in the following application. It has the virtue of being simple; it can be implemented repeatedly with different starting values; and it is quick to implement. We compare its performance to that of a probabilistic greedy algorithm. We also compare the predictions of the proposed model with those obtained from a logistic regression, and with a LP classifier [17]. Additionally, we assess the robustness of the solutions obtained by the proposed procedure in terms of predictive accuracy and the recovery of an optimal subset-conjunctive rule.

5. Rules for breast-cancer diagnosis

Approximately 12% of women will be diagnosed with breast cancer; 3.5% will die from it. It is the most common form of cancer and the second largest cause of cancer deaths among women. A breast cancer victim's chances of long-term survival increase with early detection, which depends on accurate diagnosis. But the most accurate test for breast cancer requires a surgical biopsy, which like any invasive procedures carries risks for a patient. The procedure is also expensive, time consuming, and stressful for a patient and her family. Recommending a biopsy is therefore not a trivial decision for a doctor. Still, the National Alliance of Breast Cancer Organizations reports that over 80% of biopsied breast abnormalities in the United States are found to be benign.

Physicians and computer scientists at the University of Wisconsin have developed a non-invasive diagnostic test that uses digital images of cells extracted from a patient's breast using a fine needle (the method of extracting the cells is called Fine Needle Aspiration). The cell images are used to score a tumor on 30 diagnostic measures (the means, variances and worst values on 10 cell characteristics). A separating plane, estimated using linear programming, is then used to predict if a tumor is benign or malignant. Details of the method, the measures and the predictive ability of the procedure are described in [17,25,29,30]. Here, we examine the performance of subset-conjunctive rules for predicting the malignancy of a tumor, using only seven of the 30 diagnostic measures.

The publicly available data comprise records on 569 patients examined by Dr. William H. Wolberg at the University of Wisconsin Hospitals in Madison.[2] The results of subsequent biopsies and follow-ups confirm malignant tumors in 212 cases and benign tumors in the rest. Prior analysis of the data suggests that the following seven predictor variables are especially important for classifying a tumor: (1) standard error of radius, (2) standard error of compactness, (3) worst radius, (4) worst texture, (5) worst smoothness, (6) worst concavity and (7) worst number of concave points. We examine the predictive ability of these variables in a subset-conjunctive formulation. As required by our model, we discretize the originally continuous variable; finer distinctions are possible, but we presently restrict the analysis to three levels per variable (Low, Medium, High). The cutoff points for the categories are the 25th and 75th percentiles of the variables across the 569 observations.

We use a deterministic greedy algorithm to identify disjunctive, conjunctive and subset-conjunctive classification rules. The algorithm is an adaptation of the greedy algorithm given above for maximum subset-conjunction. Let A denote the set of patients with a malignant tumor and let U denote the set of patients with a non-malignant tumor. For each attribute j, $j = 1, \ldots, m$, a solution corresponds to assigning a subset $S_j \subseteq \{1, \ldots, n_j\}$ of the n_j levels true (indicative of the disease), and its complement false (not indicative of the disease). A solution is a vector providing such an assignment for each attribute. The algorithm is identical to the greedy heuristic described earlier, except that one now has $n_j \geqslant 2$ levels for attribute j. Consequently, a single pass of the algorithm completes after $M = \sum_{j=1}^{m} n_j$ steps, examining the effect of switching each level of each attribute. As before, the algorithm terminates if there is no improvement after a pass; otherwise, the pass is repeated.

We retain the best solution across 100 runs, each run using a different random starting solution, for every subset of size k and for all of a series of relative weights assigned to the two types of classification errors. Let $w_A + w_U = 1$ be a normalization of the relative weights. We vary w_A between 0.5 and 0.95 in increments of 0.05. This restricts w_A to be at least as large as w_U; i.e., we assume that detecting a malignant tumor is at least as important as detecting a non-malignant tumor.

The best solution when $w_A = w_U = 0.5$ has an overall predictive accuracy of 94.36%, correctly classifying 535 of the 569 patients. The solution predicts a tumor to be malignant if its cell images show at least $k = 2$ of the following characteristics:

- “high” worst radius,
- “medium” or “high” worst concavity, and
- “high” worst number of concave points.

If we were to use this rule, we would want to record the worst readings on the three measures across as many replications as possible. However, one might not want to use the rule because it has a substantially better predictive accuracy for non-malignant cases (99.44% or 355 of 357 cases) than it does for malignant cases (84.9% or 180 of 212 cases). A system using this classification rule is good if one wants to keep as low as possible recommendations for a biopsy when a patient in fact has a benign tumor. But its failure to identify malignancy in over 15% of the cases might well be too high. As one increases the relative weights for the two types of error, one obtains rules that increases the percentage of correct malignancy predictions but decreases the percentage of correct benign predictions.

To examine the tradeoff between false negatives and false positives, we examine the solutions for the other values of w_A noted above. In each case, we use the greedy algorithm to generate solutions with different starting values and for each subset size between 1 and 7. We remove the dominated solutions for which both classification errors are higher than for another solution. Fig. 1 shows the non-dominated solutions. The highest predictive accuracy for malignant tumors is 100% (the leftmost point in the box in Fig. 1). It is achieved when $w_A = 0.95$ and it predicts a malignant tumor if the cell images for a patient display any two of the following characteristics: high standard error of radius, medium or high worst radius, high worst concavity and medium or high worst number of concave points. But the predictive accuracy for non-malignant tumors is 53%, and the overall predictive accuracy is 70.65%; both of the latter values are quite low. That is, if we use this subset-conjunctive rule, we will recommend all cases with a malignant tumor for a biopsy, and will also recommend 47% of non-malignant cases for a biopsy. Suppose a random sample of women were to be examined using the proposed rule. As about 12% of women in the population have breast cancer, the proposed rule will recommend nearly all of these for a biopsy, and about $0.47 \times 0.88 \times 100 \approx 41\%$ women who

[2] The data are available at the machine learning archives, UC-Irvine.

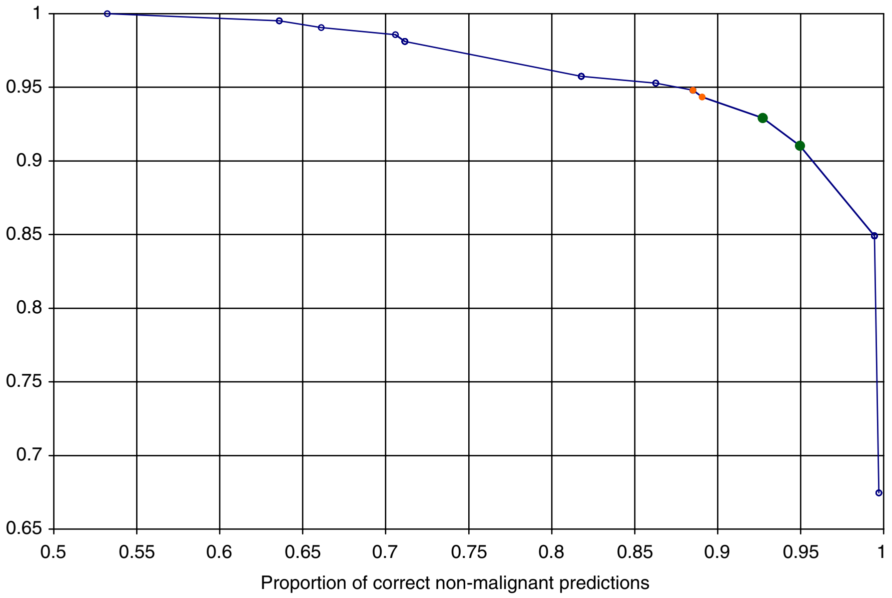

Fig. 1. Tradeoff between accuracy of malignant and non-malignant predictions.

do not have cancer for a biopsy. Thus, the final pool of biopsied women that will be correctly identified as not having cancer is about 80% of the cases $(0.41/(0.41 + 0.12) \approx 0.8)$, which is the current rate of performance for biopsies recommended by physicians. One interpretation of this is that the proposed rule does no better than the current system of clinical exams and tests. The other is that a physician can use the proposed system in place of clinical mammograms and ultrasonograms and get the same level of accuracy in the final diagnosis. Of course, using the proposed rule in conjunction with these tests and clinical exams should lead to better screening. For example, the current sample of 569 patients are consecutive patients seen by one physician. If we use the present rule to further screen the patients, we will recommend all 212 malignant cases and 167 non-malignant cases for biopsy; i.e., 44% of biopsied women will not have cancer, a substantially lower rate than the original 80%. Fig. 2 shows how the predictive accuracy for malignant and non-malignant cases varies with the subset size when $w_A = 0.95$.

How much of a tradeoff is required if we are willing to admit a subset-conjunctive rule allowing error in the detection of a malignancy? There are two solutions that give prediction accuracies over 90% for both types of error, and two others for which the error rate for non-malignant cases falls to just over 10%, pushing the correct malignant classifications to over 95%. These four solutions appear to offer the best tradeoff. Table 1 summarizes the solutions. Fig. 3 plots the two hit rates against the overall hit rate for the entire sample. We briefly discuss each of these four solutions below.

1. *Relative weights*: $w_A = 0.60,\ w_U = 0.40$: Increasing w_A from 0.5 to 0.6 introduces “high” worst texture as an additional indicator of malignancy. It also increases the number of diagnostic criteria that must be met to predict a malignant tumor as the subset size changes from two to three. The rule correctly predicts 92.9% (197/212) of the malignant cases, 92.7% (331/357) of the non-malignant cases, and 92.79% of all cases. Several different weights ranging from $w_A = 0.6$ to 0.75 identify this solution, a tumor being classified as malignant only if a patient’s profile has three or more indicators shown in Table 1 under $w_A = 0.6$.

2. *Relative weights*: $w_A = 0.65,\ w_U = 0.35$: Increasing w_A from 0.6 to 0.65 adds “medium” *s.e.* of radius and “high” *s.e.* of radius as additional factors predicting cancer; the subset size of three remains unchanged. The accuracy of predictions rises to 91.0% (193/212) for malignant tumors and drops to 93.5% (339/357) for benign tumors; overall, correct predictions decline from 94.36% to 93.5%. A tumor is predicted to be malignant only if a patient’s profile has three or more of the indicators listed under $w_A = 0.65$ in Table 1.

3. *Relative weights*: $w_A = 0.80,\ w_U = 0.20$: Increasing w_A from 0.65 to 0.8 eliminates “medium” *s.e.* of radius as a predictor of cancer, and simultaneously decreases the required value for the subset size from 3 to 2. It correctly predicts 94.8% (201/212) of the malignant cases and 88.5% (316/357) of the non-malignant cases. The overall rate of correct

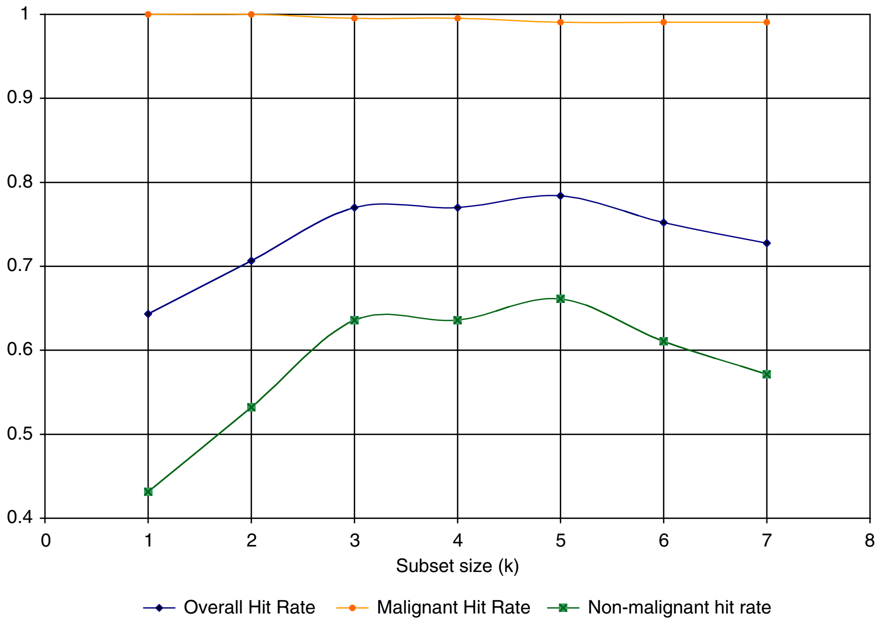

Fig. 2. Variation in predictive accuracy with subset size (k).

Table 1
Best solutions for different values of the relative weights

Variable	w_A				
	0.5	0.6	0.65	0.8	0.85
Medium *s.e.* of radius			×		×
High *s.e.* of radius			×	×	×
High worst radius	×	×	×	×	×
High worst texture		×	×	×	×
High worst smoothness					×
Medium worst concavity	×	×	×	×	×
High worst concavity	×	×	×	×	×
High number of concave points	×	×	×	×	×
Subset size (k)	2	3	3	2	3
Overall accuracy of predictions	94.36%	92.79%	91.93%	90.86%	91.05%
Accuracy of malignant predictions	84.90%	92.90%	93.50%	94.80%	94.30%
Accuracy of non-malignant predictions	99.44%	92.70%	91.00%	88.50%	89.10%

predictions is 90.86%. The rule predicts a malignant tumor only if a patient's profile has two or more of the indicators listed under $w_A = 0.80$ in Table 1.

4. *Relative weights*: $w_A = 0.85$, $w_U = 0.15$: Further increasing w_A from 0.8 to 0.85 re-introduces "medium" *s.e.* of radius as a predictor of cancer; it also reverts back to a subset size of three. The rule correctly predicts 94.3% (200/212) malignant cases; 89.1% (318/357) non-malignant cases; and 91.05% of all cases. It predicts a malignant tumor only if a patient's profile has three or more of all the indicators in Table 1.

The relation between the above sequence of successive rules is at least superficially reassuring: each is a modification of the other, rather than a completely different rule. A common set of diagnostic criteria appear across the rules: "high" worst radius, "medium" or "high" worst concavity and "high" number of concave points.

One way to assess the stability of the solutions is to estimate the rules using a randomly selected sub-sample comprising a fraction f of the data and comparing the best rules for a given weight across r runs. We use $f = 0.9$ and

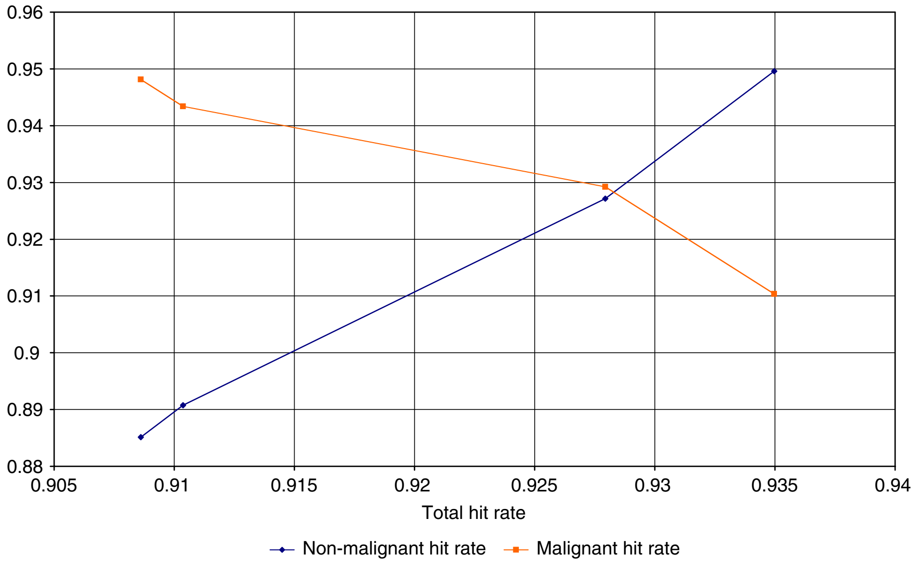

Fig. 3. Relative performance of four best diagnostic solutions.

$r = 100$ for the second solution above, (relative weights: malignant $= 0.65$, non-malignant $= 0.35$), for which the two types of classifications are nearly equal. In 99 of the 100 cases, we obtain the same solution shown under the $w_A = 0.65$ column in Table 1, which suggests a high degree of solution stability.

The error tradeoff is evidently central to the selection of a screening rule. The best rules itself is rather simple in each case, and can be readily used by a clinician as an input to deciding whether or not to send a patient for a biopsy. Such a system can realistically be used to build a clinical device that reads in the cell samples for a patient and predicts the odds of cancer for any of several possible rules a doctor might favor.

5.1. Solution using probabilistic greedy heuristic

For comparison, we estimate the rules for all seven subset sizes using a probabilistic greedy heuristic when $w_A = w_U = 1/2$. The proportionate probability rule,

$$p(G_T^{s'}, G_T^s) = \frac{G_T^{s'}}{G_T^{s'} + G_T^s}$$

does substantially worse than the greedy heuristic, even after 10,000 iterations. The reason for this appears to be that the marginal improvement in the solution value is small after just a few iterations; the probabilities are consequently close to 1/2 for almost all the iterations, which amounts to a sequence of random switches for the values of the variables. We tested the rule with several different starting values, and in each case the solution value did not exceed fifty percent of the solution value obtained using the deterministic greedy heuristic.

We obtain substantially better results using the rule

$$p(G_T^{s'}, G_T^s) = \frac{\mathrm{e}^{kG_T^{s'}}}{\mathrm{e}^{kG_T^{s'}} + \mathrm{e}^{kG_T^s}} = \frac{1}{1 + \mathrm{e}^{k(G_T^s - G_T^{s'})}},$$

with $k = 1$. For all seven subset sizes, the best solution produced by this rule over 1000 iterations is the same as the solution produced by the deterministic greedy heuristic. These limited results suggest that, depending on the problem set, the exponential form for the probabilities is better; the parameter k allows for testing different weighting schemes, and in other instances it might be useful to run the algorithm for various values of k.

5.2. Robustness

We assess the robustness of the proposed procedure by estimating the rules for $w_A = 1/2$ on a fraction *f* of the data, then using the rules to predict the classification for the remaining $1 - f$ fraction of the data. We vary the values of *f* from 0.50 to 0.90 in steps of 0.10. We perform the analysis 100 times for each value of f, randomly partitioning the data into estimation and validation samples. We record for each run the predictive accuracy (% correct classification) in the holdout and estimation samples, and when the solution (i.e., parameter estimates) obtained is different from the solution obtained using all the data. The average hit rates for correct classification for the holdout and estimation samples, and the percentage of solutions coincident with the full-data solution, are as follows:

f	Holdout sample	Estimation sample	Full-sample solutions
0.90	0.939	0.940	0.960
0.80	0.942	0.940	0.940
0.70	0.936	0.940	0.940
0.60	0.937	0.940	0.990
0.50	0.930	0.944	0.840

These results suggest that the solution we obtain is robust. The percent of correct classification is excellent for both the holdout and estimation samples (about 94% on average) and is not affected by the value of *f*. The percent of solutions that are concordant with the full-sample solution is also excellent (also about 94%) which suggests that the parameter estimates are quite stable across the different values of *f*.

5.3. Competing models

We compare the predictive performance of the proposed model to two other models: a logistic regression using the same seven predictor variables; and a LP classifier [17] that uses 30 continuous predictors in contrast to the present seven.

A stepwise logistic regression, with selection entry set at the $p = 0.05$ level, retains only main-effects terms. The estimated logistic regression equation is

$$u = 10.05 - 8.64x_{13} - 5.43x_{23} - 3.53x_{14} - 1.97x_{24} - 7.82x_{17} - 4.9x_{27};$$
$$LL = -83.84,$$

where *u* is the logit of the probability of accepting the alternative; *LL* is the log-likelihood value; x_{1j}, x_{2j}, x_{3j} denotes "low," "medium" and "high" values for variable j; and $j = 1$ refers to std. error of radius, $j = 2$ to std. error of compactness, $j = 3$ to worst radius, $j = 4$ to worst texture, $j = 5$ to worst smoothness, $j = 6$ to worst concavity and $j = 7$ to worst number of concave points. The logistic-regression and subset-conjunctive models are similar in terms of their overall fits, but can make different predictions because the former has a compensatory structure and the latter has a non-compensatory structure. For example, setting $x_{13} = x_{17} = 1$, and $x_{23} = x_{14} = x_{24} = x_{27} = 0$ in the logistic regression gives

$$u = 10.05 - 8.64 - 7.82 = -6.41; \quad p(\text{malignancy}) = \frac{e^{-6.41}}{1 + e^{-6.41}} = 0.0016.$$

That is, cell samples with "low" worst radius ($x_{13} = 1$), "low" worst number of concave points ($x_{17} = 1$) and "high" values on the other characteristics, are associated with a very small probability of malignancy. The corresponding subset-conjunctive model with $t = 2$ predicts malignancy. Unfortunately, there are no cases in the sample to test for this difference (and other similar differences) in the predictions of the two models, for the reason that cancerous cells are simultaneously altered on several cell features. We therefore cannot say that the underlying process is a subset conjunction, but only that the data are consistent with the process, and that there are conditions (albeit unobserved in the present instance) where the outcomes can differ substantially from the predictions of a logistic regression.

To use the logistic regression model for cancer diagnosis, one has to select a probability cutoff for classifying a tumor as benign or malignant. As all observations have equal weights in logistic regression (i.e., it implicitly assumes

$w_A = w_U$), we classify a tumor as malignant if the probability of malignancy exceeds 1/2. The corresponding hit rate is 94.20%, which is almost identical to the 94.36% hit rate for the subset-conjunctive rule with $w_A = w_U$. We also test the predictive validity of the logistic regression and subset-conjunctive models by running a 10-fold cross validation. We randomly select 90% of the observations for model estimation and use the remaining 10% for prediction. We repeat this analysis 100 times. For logistic regression, the mean hit rate across randomly drawn holdout samples is 93.53%, which is almost the same as the 93.90% value for the subset conjunctive model. Thus, from a predictive standpoint, the proposed model performs as well as a logistic regression, but offers two advantages over it. First, the rules obtained are simple enough to be readily used by a clinician, without having to resort to a calculation of a logit probability. Second, the availability of different, simple rules that trade off between the two types of classification errors allows a physician a choice among diagnostic criteria suiting his/her risk attitude.

The hit rates on prediction samples are slightly lower for both the logistic regression and the subset-conjunctive models than for a linear-programming classifier used by Wolberg et al. [17]. The latter has a mean hit rate of 97.5% in 10-fold cross validations. There are at least two reasons for the slight reduction in holdout performance. First, we use seven predictor variables, whereas the LP classifier uses thirty. Second, we discretize the variables; the LP classifier uses continuous predictor variables.

6. Conclusion

The forgoing results suggest that it is quite possible to get very good, simple rules for breast-cancer diagnosis that can be used in clinical settings: Fine Needle Aspiration is often performed, and it is not necessary to have sophisticated models for interpreting the data. On the other hand, there is no single dominant rule, but a collection of related rules that explicitly trade off between false negatives and false positives. This makes it possible for a physician to make explicit the assumptions in making judgments, which are otherwise implicit in clinical diagnosis of a disease. The predictive validity of the rule is impressive, comparable to that obtained in earlier models with many more predictor variables.

The model we describe is closely linked to the Logical Analysis of Data (LAD). The types of logical rules we infer permit an integer-programming formulation and the design of heuristics like LP rounding that exploit the structure of the formulation. It is not surprising that, like so many other interesting combinatorial problems, inferring subset-conjunctive rules is NP-Hard, or that it admits no constant-ratio approximation algorithm. As in the present instance, simple algorithms, like greedy heuristics, do very well in practice. We use a greedy algorithm to identify a Pareto-efficient set of subset-conjunctive rules; describe how the rules change with a re-weighting of the type-I and type-II errors; how the best rule changes with the subset size; and how much of a tradeoff between the two types of error is required by selecting a more stringent or more lax classification rule.

As noted by Alexe et al. [1], most medical literature on risk stratification focuses on specific predictors of risk. Lesser attention is devoted to interactions of risk factors. Like LAD, the subset-conjunctive model is a model of interactions. As with LAD, it is possible that the interactions revealed here may stimulate research for a better understanding of the related cause–effect relationships.

References

[1] S. Alexe, E. Blackstone, P.L. Hammer, H. Ishwaran, M.S. Lauer, C.E. Pothier Snader, Coronary risk prediction by logical analysis of data, Ann. Oper. Res. 119 (2003) 15–42.

[2] R.L. Andrews, T.C. Srinivasan, Studying consideration effects in empirical choice models using scanner panel data, J. Market. Res. 32 (1995) 30–41.

[3] E. Boros, P.L. Hammer, Pseudo-Boolean optimization, Discrete Appl. Math. 234 (2002) 155–225.

[4] E. Boros, P.L. Hammer, J.N. Hooker, Predicting cause–effect relationships from incomplete discrete observations, SIAM J. Discrete Math. 7 (1994) 531–543.

[5] E. Boros, P.L. Hammer, J.N. Hooker, Boolean regression, Ann. Oper. Res. 58 (1995) 201–226.

[6] E. Boros, P.L. Hammer, T. Ibaraki, A. Kogan, Logical analysis of numerical data, Math. Program. 79 (1997) 163–190.

[7] E. Boros, P.L. Hammer, T. Ibaraki, A. Kogan, E. Mayoraz, I. Muchnik, An implementation of logical analysis of data, IEEE Trans. Knowledge Data Eng. 12 (2) (2000) 292–306.

[8] C.H. Coombs, Mathematical models in psychological scaling, J. Amer. Statist. Assoc. 46 (256) (1951) 480–489.

[9] Y. Crama, P.L. Hammer, T. Ibaraki, Cause–effect relationships and partially defined Boolean functions, Ann. Oper. Res. 16 (1988) 299–325.

[10] R.M. Dawes, The robust beauty of improper linear models in decision making, Amer. Psychol. 34 (1979) 571–582.

[11] O. Ekin, P.L. Hammer, A. Kogan, Convexity and logical analysis of data, Theoret. Comput. Sci. 244 (2000) 95–116.

[12] D. Grether, L. Wilde, An analysis of conjunctive choice: theory and experiments, J. Consumer Res. 10 (4) (1984) 373–386.

[13] J. Huber, N. Klein, Adapting cutoffs to the choice environment: the effects of attribute correlation and reliability, J. Consumer Res. 18 (December) (1991) 346–357.

[14] R. Kohli, R. Krishnamurti, P. Mirchandani, The minimum satisfiability problem, SIAM J. Discrete Math. 7 (2) (1994) 275–283.

[15] E.K. Lee, A. Sofer, Preface, Ann. Oper. Res. 119 (2003) 13–14.

[16] D.A. Lussier, R.W. Olshavsky, Task complexity and contingent processing in brand choice, J. Consumer Res. 6 (2) (1979) 154–165.

[17] O.L. Mangasarian, W.N. Street, W.H. Wolberg, Breast cancer diagnosis and prognosis via linear programming, Oper. Res. 43 (4) (1995) 570–577.

[18] R. Motwani, P. Raghavan, Randomized Algorithms, Cambridge University Press, New York, 1995.

[19] Y. Nakajima, M. Hotta, A developmental study of cognitive processes in decision making: information searching as a function of task complexity, Psychol. Rep. 64 (February) (1989) 67–79.

[20] R.W. Olshavsky, F. Acito, An information processing probe into conjoint analysis, Decision Sci. 11 (1980) 451–470.

[21] J.W. Payne, Task complexity and contingent processing in decision making: an information search and protocol analysis, Org. Behav. Human Perform. 16 (1976) 366–387.

[22] J.W. Payne, J.R. Bettman, E.L. Johnson, Adaptive strategy selection in decision making, J. Exp. Psychol. Learn. Mem. Cogn. 14 (1988) 534–552.

[23] J. Roberts, J. Lattin, Development and testing of a model of consideration set composition, J. Market. Res. 28 (1991) 429–440.

[24] V. Srinivasan, A conjunctive-compensatory approach to the self explication of multiattributed preferences, Decision Sci. 19 (1988) 295–305.

[25] W.N. Street, W.H. Wolberg, O.L. Mangasarian, Nuclear feature extraction for breast tumor diagnosis, IS&T/SPIE 1993 International Symposium on Electronic Imaging: Science and Technology, 1905, San Jose, CA, 1993, pp. 861–870.

[26] K.H. Teigen, M. Martinussen, T. Lund, Linda versus world cup: conjunctive probabilities in three-event fictional and real-life predictions, J. Behav. Decision Making 9 (1996) 77–93.

[27] V.V. Vazirani, Approximation Algorithms, Springer, New York, 2001.

[28] M.R.M. Westenberg, P. Koele, Multi-attribute evaluation processes: methodological and conceptual issues, Acta Psychol. 87 (1994) 65–84.

[29] W.H. Wolberg, W.N. Street, D.M. Heisey, O.L. Mangasarian, Computerized breast cancer diagnosis and prognosis from fine needle aspirates, Arch. Surg. 130 (1995) 511–516.

[30] W.H. Wolberg, W.N. Street, O.L. Mangasarian, Machine learning techniques to diagnose breast cancer from fine-needle aspirates, Cancer Lett. 77 (1994) 163–171.

[31] P.L. Wright, Consumer choice strategies: simplifying versus optimizing, J. Market. Res. 11 (February) (1975) 60–67.