A Genetic Algorithm System to Find Symbolic Rules for Diagnosis of Depression

Christopher N. Chapman
Lana Deaton, Angela Harris, Nova Robinson
Department of Psychology
The University of Tulsa
600 S. College Ave.
Tulsa, OK 74104
chapman@post.harvard.edu

Abstract- A machine learning method is proposed for automatically finding psychiatric diagnostic rules. It is proposed that a genetic algorithm (GA) system can find symbolic, easily readable rules that could be used by psychiatric clinicians. Diagnosis of major depressive disorder is considered. A sample of 320 subjects with symptom information and pre-assigned diagnosis is used to train a GA model and two other statistical models, discriminant analysis and logistic regression. Each model is able correctly to classify more than 91% of cases. The GA model performs best of the three methods and yields readable, non-numeric rules.

1 Introduction

There are many different systems for diagnosis of psychiatric disorders. For example, one recent project examined 16 different systems for diagnosing schizophrenia (Herron, Schultz, & Welt, 1992). The current comprehensive psychiatric diagnostic system, the Diagnostic and Statistical Manual of Mental Disorders, 4th edition (DSM-IV), represents a consensus view of various diagnoses (American Psychiatric Association, 1994). Specific DSM-IV diagnostic rules have been developed through committee discussion and negotiation, followed by empirical trials to examine the reliability of those rules.

DSM-IV rules are presented in a logical, non-quantitative format. For example, a rule may take the form "If symptoms A and B, and either C or D, are present then diagnose condition X." This contrasts typical empirically derived rules that require calculations based on numerical weights derived from regression equations. For example, a rule could be "IF (1.2A + 0.8B + 0.45C + 0.6D – 0.15C*D) > 2.9 THEN diagnose X". Empirically derived actuarial rules have been shown to be superior to clinical judgment (Dawes, Faust, & Meehl, 1989), but are often difficult for clinicians to apply in daily practice.

Machine learning techniques may be able to find patterns in empirical case data and derive diagnostic classification rules with a logical format similar to those of DSM-IV. If so, one might be able to find rules that are as reliable and accurate as other empirically derived rules but are presented in a symbolic style that is easy for clinicians to learn and use.

The present work examines the possibility of using a genetic algorithm (GA) system to find symbolic diagnostic rules for diagnosis of depression. We compare the performance of the GA system to two classification methods commonly used in psychiatric research, discriminant function analysis (DA) and logistic regression (LR).

2 Symbolic Rules for Depression

Diagnosis of major depression in DSM-IV requires primary consideration of the following 9 symptoms: (1) depressed mood, (2) decreased interest in activities, (3) significant weight loss or gain, (4) insomnia or hypersomnia, (5) motor agitation or retardation, (6) fatigue, (7) feeling worthless or guilty, (8) poor concentration or indecisiveness, (9) thoughts of death or suicide. A major depressive "episode" is diagnosed if a patient has at 5 of these 9 symptoms, including either (1) or (2) (DSM-IV, p. 327). Finally, the diagnosis of major depressive "disorder" further requires that a patient not have a history of manic symptoms, and that the symptoms do not better fit a diagnosis of a psychotic disorder (DSM-IV, p. 344).

To use a computerized search algorithm, one must represent this kind of rule in symbolic form. A generalized scheme for rules similar to this would include the following elements: (a) a list of symptoms; (b) the possibility that individual symptoms will be (i) required, (ii) optional, or (iii) ignored; and (c) some specific number of optional symptoms that must be present. The rule for depression has 9 optional symptoms, with a required number of 5 of them. Of course, individual symptoms may be required or optional either as positive or negative symptoms (i.e., a rule might require a symptom to be either present or absent). The requirement that either symptom (1) or (2) be present from the list above is dropped because clinical experience (and combinatorics) suggests that anyone with 5 of the 9

0-7803-5536-9/99/$10.00 ©1999 IEEE

symptoms will almost always have (1) or (2).

A symbol is assigned to each symptom position in a rule to denote whether that symptom is (i) required as a positive symptom (symbol "R"), (ii) required negative ("Q"), (iii) optional positive ("O"), (iv) optional negative ("N"), or (v) ignored ("I"). A complete rule may be written by simply listing those symbols in order, appended by the number of optional rules that are needed. The rule for depression, using the 9 symptoms listed above, would be "OOOOOOOOO 5"; all 9 symptoms are optional, and at least 5 of them are needed. In practice, the scheme would be extended to consider the possible utility of other symptoms by adding more symptoms to the list. An additional value is coded at the end to specify a minimal duration of symptoms that may be required, such as one week duration of symptoms. Other conceivable rules could be constructed by changing elements of that string.

This linear format gives a data structure that is suitable for computer search techniques. Using this scheme, any rule that may be written is also decipherable into a format that would be similar in style to the rules in DSM-IV. Thus, any rule that could be found would represent a rule that is easy for clinicians to understand, learn, and use.

The scheme presented above defines a space of possible diagnostic rules. We propose that a genetic algorithm (GA) search method is an appropriate way to search that rule space. In a GA system, the population would consist of candidate rules with the structure outlined in the scheme above. A fitness function would evaluate how well each candidate diagnostic rule performs on a sample of cases with known diagnoses. The GA would run until a best rule is found, and that rule would then be evaluated on a different sample of cases. The entire model would be run repeatedly in order to obtain information about typical and best-case performance, and to use different samples of training cases. Criterion diagnoses for training and evaluation would be supplied by a procedure that assigns diagnoses using DSM-IV rules.

GA's have been successfully applied to finding diagnostic rules for female urinary incontinence using a modification of a Pittsburgh/De Jong categorical GA model (Laurikkala & Juhola, 1998; De Jong, Spears, & Gordon, 1993). However, the De Jong model does not allow all of the kinds of optional, required, positive, and negative matching rules that are desirable when finding rules similar to those in DSM-IV. In the present study, we use a modified Michigan/Holland GA model (Holland, 1992; Goldberg, 1989) that can encode such rules.

3 Performance Assessment

Performance of a diagnostic method may be assessed in various ways. When a criterion diagnosis is available, the fundamental measure of diagnostic success is the simple accuracy rate, the proportion of cases that are successfully diagnosed both positively and negatively. It is also important to consider whether this accuracy rate exceeds what might be expected by random assignment. A "chance-corrected" accuracy rate is given by Cohen's kappa statistic (Cohen, 1960). The distribution of the kappa statistic is highly dependent upon classification base rates, and thus it is not an appropriate metric for comparison of procedures (Uebersax, 1987). However, significant kappa values do establish that diagnostic agreement exceeds what may be expected from chance. In this study, we use diagnostic accuracy rate as the principal metric for comparison, and report kappa values to assess performance better than chance. The range of the kappa statistic is $(-\infty, 1.0]$; performance better than chance is shown by values greater than 0.

Two classification procedures have been selected as test models for comparing the performance of the GA model. The first method, discriminant function analysis uses a linear regression model to predict category membership from a set of predictor variables. A discriminant function takes the form $C_j = c_{j0} + c_{j1}X_1 + c_{j2}X_2 + \ldots c_{jp}X_p$, where C_j is the classification score for group j, c_j is a coefficient weight derived for variable X_j, and X_j is the raw score of each of p predictors (Tabachnick & Fidell, 1996, p. 517). The values of the discriminant function are mapped onto values of the classification variable. In the case of diagnosis, the classification value is dichotomous. The discriminant function is developed on a training set of subject data and then evaluated against a test set of data. Classification works as follows. A mean (centroid) value of the discriminant function is computed for the training cases in each category. A test case is assigned to the category whose centroid value is closest to the value of the discriminant function computed from the case's predictor values. A primary limitation of discriminant function analysis is that multivariate normality of predictor variables is assumed and the model is highly sensitive to outlying values and heterogeneous covariance among predictors.

The second method, logistic regression, finds a linear regression equation that is used in an exponential function to derive a probability of category assignment. With a dichotomous outcome measure, the probability estimate for assignment of any case i to one outcome is $Y_i = e^u / (1+e^u)$, where $u = A + B_1X_1 + B_2X_2 + \ldots B_kX_k$ for constant A, coefficient weights B_j, raw scores X_j, and k predictors (Tabachnick & Fidell, 1996, p. 576). Equation u is found by regression methods that attempt to maximize the likelihood of correct classification for the data set. With dichotomous categories, test cases are classified into the target category if their computed probability of membership is greater than 0.50, and otherwise assigned to the other category. Advantages of logistic regression are that few assumptions are made about data distribution. However, in situations where the set of predictor variables is not small, the procedure is likely to produce unstable results and may entirely fail to find solutions. As discussed below, the number of predictors used for logistic regression analysis

was limited in this study.

Performance of the GA model may be compared to performance of the two statistical models by repeatedly running each of the three procedures using different samples of training and test cases. Information can be collected on the classification performance of each run, and then standard univariate techniques may be used to compare the average performance of each method. In addition, it is important to assess typical and best-case performances, and whether methods performed better than chance.

4 Method

4.1 Participants

Subject data was gathered from archival records of inpatient admissions to a state psychiatric hospital in Oklahoma. A list was gathered of patients who had been referred for diagnostic evaluation to the facility's psychology division during the years 1996 and 1997. Subjects were required to have a full set of chart materials including discharge summary and DSM-IV diagnosis and were eliminated if the primary diagnosis was intoxication or malingering.

4.2 Symptom List

Symptoms to be coded were selected by compiling a list of the key DSM-IV symptoms for (a) major depressive episode, (b) schizophrenia, and (c) delusional disorder. This gave a total of 23 symptoms to be reviewed for each subject as shown in Table 1. In addition, information was collected on the estimated duration of the presenting symptoms.

4.3 Symptom Coding

Chart review was conducted for each subject by one of four clinical psychology graduate students. Raters were instructed to review the admission notes, psychology division reports, social work reports, and discharge summaries for each patient in order to determine what the patient's presentation was at time of admission to the hospital. Each of the 23 symptoms was rated on a four-point scale: (0) symptom definitely not present, (1) unknown, (2) present to some extent, and (3) definitely present. Raters were specifically instructed to ignore subjects' chart diagnoses when rating symptoms. After data were collected, symptom lists were recoded to dichotomous values: (0) symptom not displayed or unknown, (1) symptom displayed partially or fully.

4.4 Assignment of diagnosis

Criterion diagnoses were assigned to each subject by a computer program written by the first author to implement DSM-IV rules. Subjects were assigned a diagnosis of depression by that program if they met the DSM-IV criteria noted above. Individuals with hospital chart diagnoses of bipolar disorder were not diagnosed as depressed. This program also assigned diagnoses of schizophrenia and non-schizophrenic psychosis in accordance to DSM-IV rules, and also assigned subjects to a residual category of "other or none."

1.	Depressed mood most of the day, nearly every day
2.	Diminished interest or pleasure in activities
3.	Significant weight loss or gain
4.	Insomnia or hypersomnia, nearly every day
5.	Psychomotor agitation or retardation, nearly every day
6.	Fatigue or loss of energy, nearly every day
7.	Feelings of worthlessness or excessive guilt
8.	Indecisiveness or loss of concentration
9.	Recurrent thoughts of death or suicide
10.	Bizarre delusions (implausible or not understandable)
11.	Non-bizarre delusions (false or groundless beliefs)
12.	Preoccupation with delusions
13.	Hallucinations (any kind)
14.	Auditory hallucinations
15.	Disorganized speech (e.g., derailment, incoherence)
16.	Disorganized behavior (e.g., agitation, silliness)
17.	Catatonic behavior (e.g., bizarre postures, stupor)
18.	Negative symptom: Alogia (poverty of speech)
19.	Negative symptom: Avolition (little or no goal-directed behavior)
20.	Negative symptom: Flat affect (emotionally non-responsive)
21.	Negative symptom: Other or unspecified
22.	Social or occupational dysfunction (marked drop in achieved/expected status)
23.	Inappropriate affect (e.g., laughter with no humorous cue)

Table 1: List of Coded Symptoms
(American Psychiatric Association, 1994, pp. 285, 301, 329).

4.5 GA Model Overview

The genetic algorithm model used the following general procedure. First, the subject data set was randomly divided into approximately equal sized groups of subjects. One group was used as a training set for providing feedback to the GA model; the other group was a test set for evaluating the model's best result. Second, the GA model was initialized to random candidate rules with the genome structure described below. Third, each candidate rule was tested against the training set of subjects. If the proposed candidate rule was a positive match against a subject's symptom list, then a proposed positive diagnosis of depression was made; otherwise a negative diagnosis was proposed. The fitness function described below returned scores according to the accuracy of the candidate rule's diagnoses. Fourth, candidate rules were ranked and selected for reproduction, crossover, and mutation using the GA parameters shown below, and the evolutionary cycle was run for a fixed number of generations. Fifth, the best candidate rule was recorded and tested for performance against the test

set of subjects. Sixth, the entire GA system was run 100 times to collect data on typical and best performance. The GA system was implemented in C++ using the GAlib 2.4.2 library by M. Wall (Wall, 1996).

4.6 GA Genome Structure

The genome encoded three aspects of diagnostic rules. First, it coded 23 values for symptom matching. A scheme was used where five phenotypic values were available: "R" for absolutely required matching symptoms, "Q" for required non-matches (negative matches), "O" for optional matches, "N" for optional non-matches, and "I" for ignore. These five values were interpolated from real-valued genotype alleles in the range [0.00, 1.00], step value 0.01, and were ranked in the order, "Q," "N," "I," "O," "R".

Second, the genome coded a value for the number of optional matching symptoms required for diagnosis, in the range of [0, 23]. It was hypothesized that this value would closely co-evolve with the symptom list, and therefore the value was represented in a distributed fashion in 23 genotype alleles, each ranging [0.00, 1.00] and interspersed with the 23 symptom alleles. The value of the phenotypic optional match number was found by taking the integer floor of the sum of those alleles. Third, a value was coded for required minimal symptom duration using a five-point scale: (0) no minimal duration, (1) presence for more than one day, (2) one month or more, (3) six months or more, (4) two years or more. This was represented by 2 genotype alleles in the range [0.00, 1.00] that were summed and interpolated into five phenotypic values.

To summarize, the genome layout was as follows. The genotype included 48 real-valued alleles in the range [0.00, 1.00], step value 0.01. Positions 0, 2, 4, ... 44 represented the symptom list and were mapped to the values "Q," "N," "I," "O," and "R" noted above. Positions 1, 3, 5, ... 45 were summed and the integer floor formed the phenotypic value of the number of optional symptoms required for a match. Positions 46 & 47 were added and mapped to the five duration values noted above. The phenotype was decoded as a list of 23 symptom match indicators, plus one value for the number of optional matches and one code for duration of the disorder. This phenotype was used to match subjects' data for determination of fitness evaluation and final performance. Approximately 1.4×10^{18} unique phenotypic rules could be coded in this structure.

4.7 Fitness function

The fitness function was designed to reward correct positive diagnosis and punish incorrect positive diagnosis, in order to find rules that would maximize the chance that positive diagnoses would be correct. A given rule was matched against each case in the training set. If the subject data and candidate rule matched, then a positive diagnosis was proposed; otherwise a negative diagnosis was proposed. If a positive diagnosis was correct, a value of +1.0 was scored for that case. If incorrect, -1.0 was scored.

Partial credit was awarded as follows. If a negative diagnosis was incorrect, a "distance" score was calculated to determine how far the rule was from detecting that case. Representing D_R as the number of unmatched required symptoms (constrained to [1 ...]), and D_O as needed but unmatched optional symptoms (also [1...]), the partial score was calculated as $D = 1/D_R + 1/D_O$. This partial score was then scaled by 0.333 to fall in the range (0.00, 0.666].

Correct negative diagnoses scored nothing. The complete fitness score for a candidate rule was the sum of all true positive, false positive, and false negative (partial) scores across all training cases. This sum was scaled by the number of cases to fall in a range of [0.0, 100.0], allowing rough comparison between GA runs (note that internally the GA further performed linear scaling as noted below).

4.8 GA Parameters

A classical, generational GA model was used with the parameters shown in Table 2. These parameters were selected on the basis of review of parameters commonly used in GA applications (e.g., Goldberg, 1989).

GA type:	Classical GA model with elitism
Population size:	30
Generations run:	2000
Genome structure:	Array of real number alleles
Genome length:	48 alleles
Allele range:	[0.00, 1.00], step value 0.01
Phenotype mapping:	described in text
Selection method:	Fitness proportionate (roulette)
Fitness scaling:	Linear scaling
Elitism:	Single best genome preserved
Crossover type:	Uniform crossover
Crossover rate:	0.85 per copy
Mutation type:	Gaussian mutation
Mutation rate:	0.04 per allele
Random function:	Ran2 algorithm (Press et al, 1989)
Total runs:	100

Table 2: Genetic Algorithm Parameters

4.9 Comparison models

Discriminant function analysis (DA) was performed as follows. The data file was randomly split into training and test groups (50% chance for assignment of each case to each group). Using the training set, all 23 symptom variables and the coded duration value were entered into the regression procedure to find a classification function for the diagnostic category. This function was evaluated on the test set of subjects and accuracy statistics were recorded. The entire procedure was repeated 100 times in order to assess the typical performance across different training samples (the algorithm itself is deterministic for each sample). The procedure was performed using SPSS 8.0.1 software, and default settings were used to control model iteration,

termination, and classification.

Logistic regression analysis (LR) was performed using the same general procedure of splitting the data set, training a model, testing that model, and repeating the procedure 100 times with new training and test samples. However, early exploratory runs established that the logistic procedure was unstable when performed on sets of 23 and 14 predictor variables. For that reason, the set of predictors was limited to the following 11 items: (1) depressed mood, (2) diminished interest in activities, (3) psychomotor agitation/retardation, (4) feelings of worthlessness or guilt, (5) thoughts of death/suicide, (6) bizarre delusions, (7) non-bizarre delusions, (8) auditory hallucinations, (9) disorganized speech, (10) flat affect, and (11) social decline. Duration was not used as a predictor. These items were selected to meet criteria of clinical relevance, likelihood to be noted in clinical reports, and distribution among symptoms of depression and other disorders. The logistic regression procedure was performed with SPSS 8.0.1 software; default settings were used for model iteration, termination, and classification.

4.10 Performance measures

For each run of each model, statistics were collected on the performance of the model on the test set of data, allowing computation of the overall accuracy rate and the kappa value for diagnostic agreement. Performance of each model is assessed by examining the mean, standard deviation, median, best, worst, and 95th percentile of the overall accuracy rate of the model on 100 runs. Also, the median kappa value is reported to assess whether the model performed better than chance. Comparison between models is made using univariate analysis of variance (ANOVA) to test whether there is significant difference between models in mean accuracy performance. Pairwise comparison of models is performed using ANOVA contrasts. Normal distribution of the accuracy statistic is assessed using the Kolmogorov-Smirnov one-sample test for normality. The best rule found by each method is reported. All statistical significance testing is reported at the alpha level of $p < .05$.

5 Results

Demographic information about the full sample is reported in Table 3, and prevalence of computer-assigned diagnoses is shown in Table 4. Data was collected from 320 subjects and 19.1% of that full sample was given a criterion diagnosis of major depression.

As shown in Table 5, all three classification methods performed well on the diagnostic task, achieving better than 91% accuracy in most runs. In the median, 95th percentile, and best cases, the GA method appears to have performed better than the statistical methods. The GA method achieved 100% accuracy rate on test samples during 3 of the 100 runs.

Gender	N	%
Male	200	62.5
Female	120	37.5

Ethnicity	N	%
Caucasian	236	73.8
African-American	45	14.1
Native American	36	11.3
Hispanic	1	0.3
Asian American	2	0.6

Table 3: Participant Demographics

Assigned Diagnosis	N	%
Depression	61	19.1
Schizophrenia	122	38.1
Psychosis, other	45	14.1
Other or none	112	35.0

Table 4: Prevalence of Diagnoses in Full Sample
N=320. Columns do not total to 320 or 100% because some subjects had two diagnoses.

Method	Median	95th %	Best	Worst
GA	0.955	0.994	1.000	0.856
DA	0.933	0.965	0.980	0.890
LR	0.916	0.947	0.970	0.854

Table 5: Typical Accuracy Rate by Classification Method
Each model was run 100 times.

Distribution of the accuracy rate was not significantly different from a normal (Gaussian) distribution for any of the methods (maximum Kolmogorov-Smirnov $Z = 1.234$, $N = 100$, $p > .05$). Therefore, it is feasible to use standard univariate inferential statistical tests to compare the mean performance of different methods. Rules found by the GA method achieved an average 94.8% accuracy in diagnosis, compared with 93.4% in discriminant analysis, and 91.6% with logistic regression (Table 6). The superiority of the GA method is small in absolute terms, but the difference is statistically significant on ANOVA test, $p < .05$. The kappa values show that each method performed better than chance.

Method	Mean	Std. dev.	Median Kappa
GA	0.948*	0.0378	0.851*
DA	0.934	0.0191	0.800*
LR	0.915	0.0218	0.727*

Table 6: Accuracy Rate and Reliability. N=100 runs each.
* for *mean* indicates best performer, $p < .05$.
* for *kappa* indicates performance better than chance, $p < .05$.

The best rules found by the GA procedure are shown in Table 7, along with a sample translation of the first rule. Each of those 3 rules achieved 100% accuracy in the test and training samples. The best results of discriminant analysis and logistic regression from the 100 runs of those models are shown in Tables 8 and 9.

OOOOOOOOONIIIINIQOQINIO, match 8
OOOOOOOOOIIIIIINNONINII, match 9
OOOOOOOOONIIIINNQINIIIO, match 9

No minimum duration was required by any of these rules.

The first rule above may be translated as follows:
A. For a diagnosis of major depression, 8 or more of the following signs must be present:
(1, O) Depressed mood most of the day ...
(2, O) Diminished interest or pleasure in activities
(3, O) Significant weight loss or gain
(4, O) Insomnia or hypersomnia, nearly every day
(5, O) Psychomotor agitation or retardation ...
(6, O) Fatigue or loss of energy, nearly every day
(7, O) Feelings of worthlessness or excessive guilt
(8, O) Indecisiveness or loss of concentration
(9, O) Recurrent thoughts of death or suicide
(10, N) ABSENCE of Bizarre delusions
(15, N) ABSENCE of Disorganized speech
(18, O) Negative symptom: Alogia ...
(21, N) ABSENCE of Negative symptom: Other ...
(23, O) Inappropriate affect
B. Neither of the following may be present:
(17, Q) Catatonic behavior
(19, Q) Negative symptom: Avolition

Table 7: Best Symbolic Rules Found by GA

$C = .0337\ S_1 + .6203\ S_2 + .2601\ S_3 + .2206\ S_4 + .1631\ S_5 + .4718\ S_6 + .4744\ S_7 + .3164\ S_8 + .1050\ S_9 + .1246\ S_{10} - .1797\ S_{11} - .0334\ S_{12} - .0843\ S_{13} + .0748\ S_{14} + .0296\ S_{15} - .1744\ S_{16} + .0409\ S_{17} + .6717\ S_{18} + .0917\ S_{19} + .0437\ S_{20} - .2417\ S_{21} - .0054\ S_{22} + .0317\ S_{23} + .0547$ DURATION.

Centroids: depressed = 3.306; non-depressed = -0.656

Table 8: Best Discriminant Function
S_j is raw score of dichotomous symptom variable listed in Table 1. DURATION is code for duration of illness.

$u = -13.195 + 8.466\ S_1 + 2.881\ S_2 + 1.693\ S_5 + 1.889\ S_7 + 1.964\ S_9 + .029\ S_{10} - .978\ S_{11} + .1572\ S_{14} - 1.111\ S_{15} - .079\ S_{20} + .045\ S_{22}$.

Table 9: Best Logistic Regression Function (exponent)
S_j is raw score of symptom variable listed in Table 1.

6 Discussion

The results show that each of the three classification methods performed well in finding diagnostic rules for depression in the present data set. The genetic algorithm procedure performed modestly better than the other two models in overall accuracy.

The true superiority of the GA procedure for clinicians, however, is demonstrated in Tables 7, 8, and 9. Rules found by the GA model can be precisely decoded into diagnostic rules that are relatively easy to comprehend, and are structurally similar the kind of rule that clinicians use in the DSM-IV diagnostic manual. Rules found by the regression procedures are not amenable to precise translation into non-numerical form and are more difficult to read and remember.

Because this study is an initial investigation into the utility of machine learning methods for finding symbolic diagnostic rules, there are significant limitations and many unanswered questions with the method and findings here.

First, the diagnostic problem considered was selected principally for its amenability to representation in a minimal GA model. Other clinical diagnoses require larger sets of possible rule elements and this yield a larger search space that is more difficult to search. Second, the indicator variables used here were selected because of their importance in DSM-IV. It would be more informative to consider a larger range of variables in order possibly to revise or DSM rules. Third, criterion diagnoses were assigned on the basis of DSM-IV criteria. It would be important in future studies to consider other criterion sources, such as patient self-report, expert consensus, or physical laboratory findings.

Fourth, it would be interesting to extend the rule system beyond dichotomous categories, i.e., to find diagnostic rules for multiple disorders simultaneously. Fifth, it is possible that changes to the GA system would result in better or worse performance. Elements such as other genotype-to-phenotype maps and different GA parameters and models have not been considered. Sixth, other statistical or stochastic methods might perform better than the models selected here and outperform the GA system. Seventh and finally, the population sampled in this study was limited and not necessarily representative of other psychiatric populations.

With some of these concerns in mind, a larger study is underway to investigate performance of five methods (genetic algorithm, discriminant analysis, logistic regression, simulated annealing, and random mutation search) on four psychiatric diagnoses (depression, schizophrenia, non-schizophrenic psychosis, and other).

7 Conclusion

A genetic algorithm system has been successfully demonstrated that can derive symbolic, readable diagnostic rules for major depression from a limited trial sample of cases and predictor data. This system classifies cases as reliably as discriminant function analysis and logistic regression. This positive result suggests the utility of further research on the application of genetic algorithm and other machine learning methods to psychiatric diagnosis.

Acknowledgements

The present work derives from research performed in the course of the first author's dissertation project at the University of Tulsa, Department of Psychology, 1999. That work was supervised by Brent W. Roberts. Early review

and assistance were provided by Robert Nicholson and Michael Basso of the University of Tulsa and Jeanne Russell of the University of Oklahoma School of Medicine.

Bibliography

American Psychiatric Association (1994). Diagnostic and Statistical Manual of Mental Disorders: DSM-IV (4th ed.). Washington, DC: American Psychiatric Association.

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20, 37-46.

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgement. Science, 243, 1668-1674.

De Jong, K. A., Spears, W. A., & Gordon, D. F. (1993). Using genetic algorithms for concept learning. Machine Learning, 13(2-3), 161-188.

Goldberg, D. E. (1989). Genetic algorithms in search, optimization, and machine learning. Reading, Mass.: Addison-Wesley.

Herron, W. G., Schultz, C. L., & Welt, A. G. (1992). A comparison of 16 systems to diagnose schizophrenia. Journal of Clinical Psychology, 48(6), 711-721.

Holland, J. H. (1992). Adaptation in Natural and Artificial Systems. Cambridge, MA: MIT Press.

Laurikkala, J., & Juhola, M. (1998). A genetic-based machine learning system to discover the diagnostic rules for female urinary incontinence. Computer Methods and Programs in Biomedicine, 55, 217-228.

Press, W. H., Flannery, B. P., Teukolsky, S. A., & Vetterling, W. T. (1989). Numerical Recipes in Pascal. Cambridge: Cambridge University Press.

Tabachnick, B. G., & Fidell, L. S. (1996). Using Multivariate Statistics. New York: HarperCollins.

Uebersax, J. S. (1987). Diversity of decision-making models and the measure of interrater agreement. Psychological Bulletin, 101(1), 140-146.

Wall, M. (1996). GAlib (Version 2.4.2). Cambridge, MA: MIT. URL: http://lancet.mit.edu/ga.