Generating Models of Mental Retardation from Data with Machine Learning

Subramani Mani * Suzanne W. McDermott † Michael J. Pazzani *

Abstract

This study focused on generating simple and expressive domain models of Mental Retardation (MR) from data using Knowledge Discovery and Datamining (KDD) methods. 2137 cases (mild or borderline MR) and 2165 controls (randomly selected) from the National Collaborative Perinatal Project (NCPP), a multicentric study involving pregnant mothers and the outcomes, constituted our sample. Twenty attributes (prenatal, perinatal and postnatal), thought to play a role in MR were utilized. The outcome variable (class), was, whether the child was retarded or not, based on the IQ score. Tree learners (C4.5,CART), rule inducers (C4.5Rules,FOCL) and a reference classifier (Naive Bayes) were the machine learning algorithms used for model building. The predictive accuracy ranged from 68.4% (FOCL) to 70.3% (Naive Bayes). CART obtained a sensitivity of 79.0% and also generated highly stable and simple trees across fifty random two-third (training), one-third (testing) partitions of the sample. The algorithms identified emotional/behavioral problem *in children as a significant predictor of MR risk. Our study shows that KDD methods hold promise in recovering useful structure from medical data.*

1 Introduction

Knowledge Discovery and Datamining (**KDD**) methods are emerging as useful tools to learn structure from data. They differ from traditional methods by its well defined goal of elucidating domain models such as decision trees, rules, graphs etc. from data. The KDD process involves many steps encompassing data preprocessing (attribute selection, recoding etc.), choice of datamining algorithms, running protocol and postprocessing of the output. A simple KDD model constitutes an input of processed data to the datamining engine (one or more pre-selected algorithms) and the output is "knowledge" in the form of an understandable structural model of the input data. See [7] for a detailed discussion on this. Some recent applications in the medical domain include differential diagnosis of abdominal pain [19], a screening model for dementia [24], and learning from a database of sports injuries [11].

*(mani,pazzani)@ics.uci.edu Department of Information and Computer Science, University of California at Irvine, Irvine CA 92697.

†smcdermott@fpgw.rmh.edu Department of Family and Preventive Medicine, University of South Carolina School of Medicine, Six Richland Medical Park, Columbia SC 29203.

We have selected the domain of Mental Retardation (**MR**) for our study due to its complex nature and lack of simple and insightful models. MR has a complex etiology with an interplay of genetic and environmental factors, but the causal mechanisms are not clearly understood. Mild Mental Retardation (**MMR**), with an IQ range of 50–69, constitutes eighty percent of all MR and has no known biological cause in more than half of the cases [1], [8]. In contrast, Severe Mental Retardation (**SMR**), with an IQ < 50, has a known organic basis (e.g. Down's Syndrome, anencephaly, cretinism etc.) in most instances. In this study, we have focussed on learning simple models of MMR, employing KDD methods. We believe that these models will help in early detection and intervention in MMR.

1.1 MR models from literature

The recent models discussed in literature have either a narrow focus or a very complex structure. There are conceptual models just to define MR [4], [9]. MR-Expert [10] is a rule based expert system for managing psychiatric problems in retarded individuals. There are also models linking demographic factors and prevalence rates [18], and also a linear model accounting for the variation in school district rate of MR [17]. MENTOR [14] is a complex Bayesian Network model partially learned from data. It attempts to bring out all the relationships and interactions among the variables used to build the model.

2 Description of the data set

The dataset used for this study is from the National Collaborative Perinatal Project (**NCPP**), of the National Institute of Neurological and Communicative Disorders and Stroke. It is a large longitudinal multicentric study of pregnancy outcomes, consisting of data on more than 50,000 pregnancies, between 1959 and 1974. Live births were followed up for eight years.

We identified twenty variables (prenatal, perinatal and postnatal) which are thought to play a role in MR. Since our goal was generating models for early detection and intervention, we included children in the IQ range 70–84 (> 2 SD and < 1 SD) also. This category was previously referred to as borderline MR but dropped subsequently to restrict eligibility of services (e.g. special schooling) to children with IQ below 70. Our sample included this group for two reasons— (1) to obtain a larger sample of cases and (2) the inclusion can be justified, as borderline MR is a continuum of MMR into a region of milder impairment. On the other hand, SMR has been mostly shown to be organic

in origin and probably results from cognitive damage [13], [26].

To keep the sample representative, only one record per family was selected. We identified 3598 cases who were retarded. We excluded cases with missing values resulting in 2138 cases. An equal number of controls (IQ $\geq$ 85) were also randomly selected for this study. Table 1 gives the sample characteristics for the cases and controls. We also use the terminology *retarded* and *normal* instead of *cases* and *controls*.

3 Methods

Since our goal was to develop simple and expressive domain models of MR, we selected Tree builders and Rule learners for our task. These models have the advantage that they can easily be taken *offline*, and tend to be simple and understandable models, compared with complex models such as neural networks, Bayesian networks or multiple models.

3.1 Machine Learning Algorithms

We concentrated on decision tree learners, rule learners and the naive Bayesian classifier. Decision trees and rules generate clear descriptions of how the ML method arrives at a particular classification. The naive Bayesian classifier was selected as a reference baseline classifier for comparison purposes.

MLC++ (Machine Learning in C++) [12] is a software tool developed at Stanford University which has implemented the commonly used machine learning algorithms. It also provides standardized methods of running experiments using these algorithms. C4.5 is a decision tree generator and C4.5Rules produces *if ... then* rules from the decision tree [22]. Naive Bayes is a classifier based on Bayes Rule. Even though it makes the assumption that the attributes are conditionally independent of each other given the class, it is a robust classifier and serves as a good comparison in terms of accuracy for evaluating other algorithms [6].

FOCL [20] is a concept learner which can incorporate a user provided knowledge of two types. First, when provided with a guideline or protocol directly, FOCL has the capacity for revision if the guidelines produce better classification rules than that produced from exploration of the data. Second, FOCL can accept information on each nominal variable indicating which values of the variable increase the probability of belonging to a class (such as impaired) and information on each continuous variable on whether higher or lower values of the variable increases the probability of belonging to a class. When this facility of FOCL is used, it is termed "constrained" FOCL. For this study we used only the "unconstrained" functionality of FOCL.

CART [2] is a classifier which uses a tree-growing algorithm that minimizes the standard error of the classification accuracy based on a particular tree-growing method applied to a series of training subsamples. We used Caruana and Buntine's implementation of CART, (the "IND" package) [3]. For each training set, CART builds a classification tree where the size of the tree is chosen based on cross-validation accuracy on this training set. The accuracy of the chosen tree is then evaluated on the unseen test set.

3.2 Model Building

We used MLC++ to run these ML algorithms on our dataset. The whole sample was divided randomly into two-thirds (training) and one-third (test) sets. Models were generated from the training set and evaluated on the unseen test set. This procedure was repeated fifty times. The classification accuracies reported are the mean scores obtained with the test sets.

4 Results

Table 2 gives the detailed results obtained with the various ML algorithms. The performance of the different algorithms were comparable, with only a small variation in total accuracy scores. We examined the sensitivity (probability of correctly classifying cases, i.e. the retarded children in our sample) and specificity (probability of correctly classifying controls, i.e. the normal children) for each ML run of the testing samples. CART and naive Bayes gave higher sensitivity (eight to ten percentage points more than specificity), whereas, C4.5 and C4.5Rules had almost the same sensitivity and specificity. CART gave typically simple decision trees with four to six leaves. See Figure 1 for a representative CART tree. C4.5 gave much more complex trees with more than hundred leaves in almost all the runs. The rules generated were comparatively simpler. See Figure 2 for a typical set of rules.

Table 2: **Sensitivity and Specificity of the machine learning algorithms**

(MR n = 2137, Normal n = 2165)

Algorithm	Accuracy	Sensitivity	Specificity
C4.5	68.5	68.7	68.3
C4.5Rules	68.9	69.6	68.2
Naive Bayes	70.3	75.1	65.5
CART	69.8	79.0	60.7
FOCL†	68.4	72.1	64.8

† Results from 20 runs

5 Discussion

Naive Bayes gave the highest classification accuracy but the performance of the other algorithms were comparable. CART had the highest sensitivity of all (79.0 percent), followed by naive Bayes (75.1 percent), making them preferable as screening tools for assessing MR risk. Actually, with Naive Bayes, it is straightforward to gain sensitivity or specificity depending on the requirement, by tuning the threshold probability for classification. But, Naive Bayes gives a probability density function and not a decision tree or a rule set.

CART gave simple decision trees compared to C4.5. On most runs CART gave the same decision tree, making it a very stable domain model. The most informative attribute was *presence or absence of an emotional (behavioral) problem*, which consistently occupied the

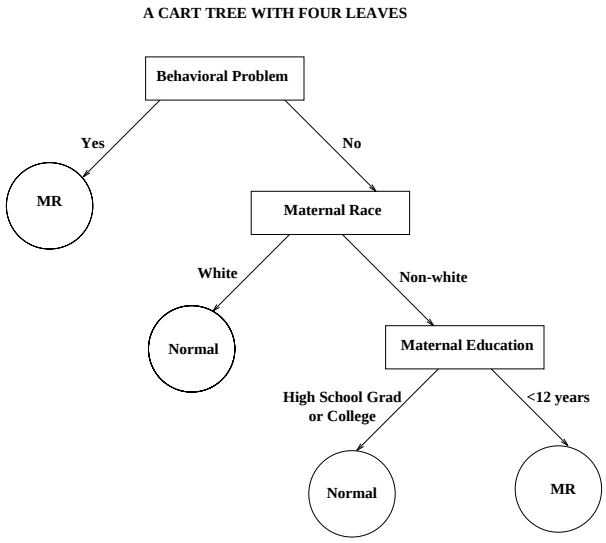

Figure 1: **Cart tree**

root of CART and C4.5 trees. The second most significant attribute turned out to be maternal race. Maternal education, birth weight of the child, maternal IQ and family income also were shown to be important by the generated models.

5.1 Model Selection

The three important considerations in model selection are its accuracy, comprehensibility and stability. There is no single formula applicable to all models. The selection criteria might vary depending on the domain and also the task the model is expected to perform. We argue that in medical domains comprehensibility is particularly important. Let us analyze these components in some more detail.

Accuracy, as is well established, cannot just be represented by one number, even though for some models just a total accuracy score might suffice. Generally at least two types of information are sought— sensitivity (predicted true positives divided by the total number of true positives) and specificity (predicted true negatives divided by the total number of true negatives). For instance, a model used for screening will lay more emphasis on sensitivity and might trade off specificity in the process. Accuracies above a threshold level chosen by the user, is a necessary pre-requisite for any model to be acceptable.

Comprehensibility of the model is of prime significance in many domains. This property of a model makes it adaptable and easily accessible for revision. A comprehensible model could be compared to a well-structured and nicely documented piece of code. Though there is some consensus on simpler (smaller) models being more comprehensible, researchers have recently focussed on factors such as adherence to domain constraints [21]. In this era of evidence-based medicine, with an attempt at making the practice of medicine much more uniform and objective, formulating effective protocols and guidelines is an active area

Figure 2: **A C4.5Rule Set**

Rule 1: birth-weight is normal *and* maternal education is college *and* maternal race is white ⇒ class **normal**(92%)

Rule 2: maternal IQ is normal *and* mother not a smoker *and* no maternal disease *and* gestation normal *and* father's education is ≤12 years *and* child is female *and* child has no emotional problem *and* no Caesarian section *and* no maternal X-ray ⇒ class **normal**(84%)

Rule 3: maternal education is at least high school grad *and* gestation not premature *and* father's education is at least high school grad *and* child has no emotional problem ⇒ class **normal**(82%)

Rule 4: maternal education less than college *and* child has emotional problem ⇒ class **retarded**(80%)

Note: The percentage figures in paretheses alongside each rule are the accuracy figures of the rule when it is applicable.

of research. For any guideline to be useful, comprehensibility is a necessary condition. There is preliminary evidence that models that do not violate the existing discipline rules are preferred by care givers [21]. In our runs, the CART models scored high on comprehensibility.

Stability of a model could be looked at from a structural and functional point of view. Higher stability means less variance in the models generated from different partitions of the sample. Structurally exact models given by a majority of random training splits of the data is a good indicator of the stability of the model. For example, the Cart tree in Figure 1 was generated by more than forty of the fifty random train sets. C4.5 trees were generally very complex and varied structurally in the different runs. But, the decision nodes of the Cart tree also figured consistently in the top region of the C4.5 trees. Highly correlated attributes can act as confounders of structural stability. Feature selection could be an option to filter out such attributes. The functional stability of structurally different models could be assessed by the variability in the classification of a representative test sample set apart for this task. Turney [25] discusses syntactic and semantic similarity measures (which correspond to structural and functional stability) for characterizing the stability of models. He has argued for semantic similarity measures for ascertaining stability since they are less sensitive to superficial variations in representations. Likewise, they could prove useful in comparisons of stability across much varied representations such as neural networks and decision trees. The reader is referred to [25] for a detailed discussion

of model stability issues.

Adequacy of a model could be extrapolated as a function of its accuracy, comprehensibility and stability. It is mostly domain and user dependent. It could be reasonably argued that in the field of medicine, for a model to be effective and useful as a guideline, it has to score high on all these components. We now move on to a short description of the salient features of our MR model from a medical viewpoint.

5.2 Highlights of our MR model

The role of parental education, particularly maternal [5], low birth weight as a risk factor for MR [16], and the role of of socio-economic status (**SES**) factors like family income (see [26] for a short discussion) have been well documented in literature.

Though researchers have discussed the association between MR and behavioral problems in children [15],[23], ML methods have identified a potential infantile marker—*emotional or behavioral problem*, as a predictor of MR risk.

In our dataset, emotional problems were ascertained from care givers at the age of four, an age when it is possible to administer IQ tests. Nevertheless, this is very much significant since there is no universal administration of IQ tests. There are also questions regarding the desirability of such a measure (apart from issues of feasibility and expenses), from a sociological perspective. But identification of behavioral problems in infants and pre-school children will facilitate proper work up and early intervention. Without getting sidetracked by the controversy of whether MMR is curable or not (the present consensus is not in favor of a cure), early intervention definitely helps children at risk to cope better in school [26],[13].

5.3 Limitations

The dataset which we used for our model building task came from a multi-centric study and hence may not be representative of the general US population. One notable feature is the over representation of non-white mothers in the sample. The enrollment period of the study was from 1959 to 1974, with an eight year follow up of live births. Some of the risk factors operable then might not have the same validity today. For example, the generated tree models feature maternal education alone (most CART trees), and the father's education node (when present) was just below the mother's. This might correctly reflect the role of mother as the primary care giver a couple of decades ago, but might not to the same extent, in the present era.

6 Conclusions

In this study, we have shown that ML algorithms can be employed to construct simple and expressive domain models of MR. Specifically, by using easily ascertainable prenatal, perinatal and postnatal parameters, (and using the Child's IQ score cutoffs as outcome) we have demonstrated the ability to assess the risk of MR with a high degree of accuracy. In the process we have succeeded in identifying a single attribute i.e. behavioral problem in infancy, as a potential predictor of MR risk.

References

[1] M.L. Batshaw. *Mental Retardation*, volume 40 of *Pediatric Clinics of North America—The Child With Developmental Disabilities*, pages 507–522. W.B.Saunders Company, Philadelphia, PA, 1993.

[2] L Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone. *Classification and Regression Trees*. Wadsworth, Belmont, 1984.

[3] Wray Buntine and Rich Caruana. *Introduction to IND Version 2.1 and Recursive Partitioning*. NASA, 1992.

[4] L. St. Claire. A multidimensional model of mental retardation: Impairment, subnormal behavior, role failures, and socially constructed retardation. *Acta Paediatrica*, 94:88–96, 1989.

[5] Carolyn D. Drews, Catherine C. Murphy, Marshalyn Yeargin-Allsopp, and Pierre Decoufle. The relationship between idiopathic mental retardation and maternal smoking during pregnancy. *Pediatrics*, 97(4):547–553, 1996.

[6] RO Duda and PE Hart. *Pattern Classification and Scene Analysis*. John Wiley, New York, 1973.

[7] Usama M. Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From data mining to knowledge discovery: An overview. In Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pages 1–36. AAAI Press, Menlo Park, California 94025, 1996.

[8] E. Fernell. Mild mental retardation in school children in a Swedish suburban municipality: prevalence and diagnostic aspects. *Acta Paediatrica*, 85(5):584–8, 1996.

[9] S. Greenspan and J.M. Granfield. Reconsidering the construct of mental retardation: Implications of a model of social competence. *American Journal on Mental Retardation*, 96(4):442–453, 1992.

[10] M.G. Hile, D.M. Campbell, B.B. Ghobary, and M.N. Desrochers. Development of knowledge bases and the reliability of a decision support system for behavioral treatment consultation for persons with mental retardation: The Mental Retardation-Expert. *Behavior Research Methods, Instruments, and Computers*, 25:195–198, 1993.

[11] I.Zelic, I.Kononenko, N.Lavrac, and V.Vuga. Machine learning applied to diagnosis of sport injuries. In Elpida Keravnou, Catherine Garbay, Robert Baud, and Jeremy Wyatt, editors, *Lecture Notes in Artificial Intelligence: Artificial Intelligence in Medicine, AIME97*, volume 1211, pages 138–144. Springer, 1997.

[12] R Kohavi, George John, Richard Long, David Manley, and Karl Pfleger. MLC++: A machine learning library in C++. In *Tools with Artificial Intelligence*, pages 740–743. IEEE Computer Society Press, 1994. Available by anonymous ftp from: `starry.stanford.edu: pub/ronnyk/mlc/toolsmlc.ps`.

[13] Donald L. Macmillan, Gary N. Siperstein, and Frank M. Gresham. A challenge to the viability of mild mental retardation as a diagnostic category. *Exceptional Children*, 62(4):356–371, 1996.

[14] Subramani Mani, Suzanne McDermott, and Marco Valtorta. MENTOR: A bayesian model for prediction of mental retardation in newborns. *Research in Developmental Disabilities*, 1997. In Press.

[15] Suzanne McDermott, Ann L. Coker, Subramani Mani, Shanthi Krishnaswami, Richard J. Nagle, Laura L. Barnett-Queen, and Donald F. Wuori. A population-based analysis of behavior problems in children with cerebral palsy. *Journal of Pediatric Psychology*, 21(3):447–463, 1996.

[16] Suzanne McDermott, Ann L. Coker, and R.E. McKeown. Low birthweight and risk of mild mental retardation by ages 5 and 9 to 11. *Pediatric and Perinatal Epidemiology*, 7:195–204, 1993.

[17] S.W. McDermott. An exploratory model to describe school district prevalence of mental retardation and learning disabilities. *American Journal of Mental Retardation*, 99:175–185, 1994.

[18] S.W. McDermott and J.M. Altekruse. Dynamic model for preventing mental retardation in the population: The importance of poverty and deprivation. *Research in Developmental Disabilities*, 15:49–65, 1994.

[19] C Ohmann, Q Yang, V Moustakis, K Lang, and van PJ Elk. Machine learning techniques applied to the diagnosis of acute abdominal pain. In Pedro Barahona and Mario Stefanelli, editors, *Lecture Notes in Artificial Intelligence: Artificial Intelligence in Medicine, AIME95*, volume 934, pages 276–281. Springer, 1995.

[20] Michael Pazzani and Dennis Kibler. The utility of knowledge in inductive learning. *Machine Learning*, (9):57–94, 1992.

[21] Michael J. Pazzani, Subramani Mani, and W.R. Shankle. Beyond concise and colorful: Learning intelligible rules. In *The third international conference on Knowledge Discovery and Datamining*, pages 235–238. AAAI Press, Menlo Park, California., 1997.

[22] JR Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, Los Altos, California, 1993.

[23] Michael Rutter, Emily Simonoff, and Robert Plomin. Genetic influences on mild mental retardation: Concepts, findings and research implications. *Journal of Biosocial Science*, 28:509–526, 1996.

[24] WR Shankle, S Mani, M Pazzani, and P Smyth. Detecting very early stages of dementia from normal aging with machine learning methods. In Elpida Keravnou, Catherine Garbay, Robert Baud, and Jeremy Wyatt, editors, *Lecture Notes in Artificial Intelligence: Artificial Intelligence in Medicine, AIME97*, volume 1211, pages 73–85. Springer, 1997.

[25] P. Turney. Bias and the quantification of stability. *Machine Learning*, 20:23–33, 1995.

[26] Edward Zigler. Editorial: Can we *cure* mild mental retardation among individuals in the lower socioeconomic stratum? *American Journal of Public Health*, 85(3):302–304, 1995.

Table 1: **Characteristics of the NCPP Sample of this study**

	Case Children (n = 2137)		Control Children (n = 2165)	
	No.	%	No.	%
		Children		
Birth weight				
Normal birth weight	1828	85.5	1983	91.6
Low birth weight	309	14.5	182	8.4
Gender				
Female	965	45.2	1149	53.1
Male	1172	54.8	1016	46.9
Health Problem				
None	648	30.3	1049	48.5
Physical	646	30.2	891	41.2
Emotional	328	15.3	99	4.6
Both	515	24.2	126	5.7
Head Circumference				
Normal	1728	80.9	1778	82.1
Abnormal	409	19.1	387	17.9
Fetal Distress				
No	1705	79.8	1615	74.6
Yes	432	20.2	550	25.4
Resuscitation				
No	2023	94.7	2022	93.4
Yes	114	5.3	143	6.6
		Mothers/Fathers		
Maternal Education				
$\leq$12 years	1489	69.7	1046	48.3
High School Grad	573	26.8	800	37.0
College	75	3.5	319	14.7
Maternal Race				
Non-white	1586	74.2	940	43.4
White	551	25.8	1225	56.6
Maternal IQ				
Normal	1666	78.0	1945	89.8
Retarded	471	22.0	220	10.2
Maternal Smoking				
No	1207	56.5	1173	54.2
Yes	930	43.5	992	45.8
Maternal Disease				
No disease	1270	59.4	1299	60.0
Disease present	867	40.6	866	40.0
Gestation				
Full term	1436	67.2	1645	76.0
Pre-mature	510	23.9	327	15.1
Post-mature	191	8.9	193	8.9
Paternal Education				
$\leq$12 years	1468	68.7	1061	49.0
High School Grad	573	26.8	703	32.5
College	96	4.5	401	18.5
Family Income				
Satisfactory	67	3.1	249	11.5
Low	2070	96.9	1916	88.5
Previous Stillbirth				
None	1871	87.6	1917	88.5
Yes	266	12.4	248	11.5
Caesarian Section				
No	2016	94.3	2051	94.7
Yes	121	5.7	114	5.3
Induced Labor				
No	2021	94.6	2021	93.3
Yes	116	5.4	144	6.7
Maternal X-ray				
No	1628	76.2	1543	71.3
Yes	509	23.8	622	28.7
Prenatal Care				
Yes	1766	82.6	1933	89.3
No	371	17.4	232	10.7
Maternal Age at Birth				
14-19	634	29.7	443	20.5
20-34	1339	62.7	1518	70.1
$\geq$35	164	7.6	204	9.4