Ensembling for Confidence: Predicting Analyte Values from Other Same-Specimen Analyte Values in the Clinical Laboratory

James M. DeLeo
Alan T. Remaley
Stephen J. Rosenfeld
National Institutes of Health Clinical Center
Bethesda, Maryland U.S.A.

presented at the
7th Course of the International School on Neural Networks
"Ensemble Methods for Learning Machines"
International Institute for Advanced Scientific Studies
Salerno, Italy

September 22 to 28, 2002

The National Institutes of Health Clinical Center

The National Institutes of Health Clinical Center, formally known as the Warren Grant Magnuson Clinical Center, is a U.S. Government funded biomedical research facility, that is part of the National Institutes of Health (NIH) in Bethesda, Maryland. As the research hospital for the NIH, the Clinical Center supports clinical investigations conducted by the Institutes. It was specifically designed to bring patient care facilities close to the research labs so that biomedical advances can be moved quickly from labs to patient treatment.

Scientific Computing Section

Jim DeLeo, Chief

Mission: "Dedicated to advancing medical and computing knowledge to improve health care for all."

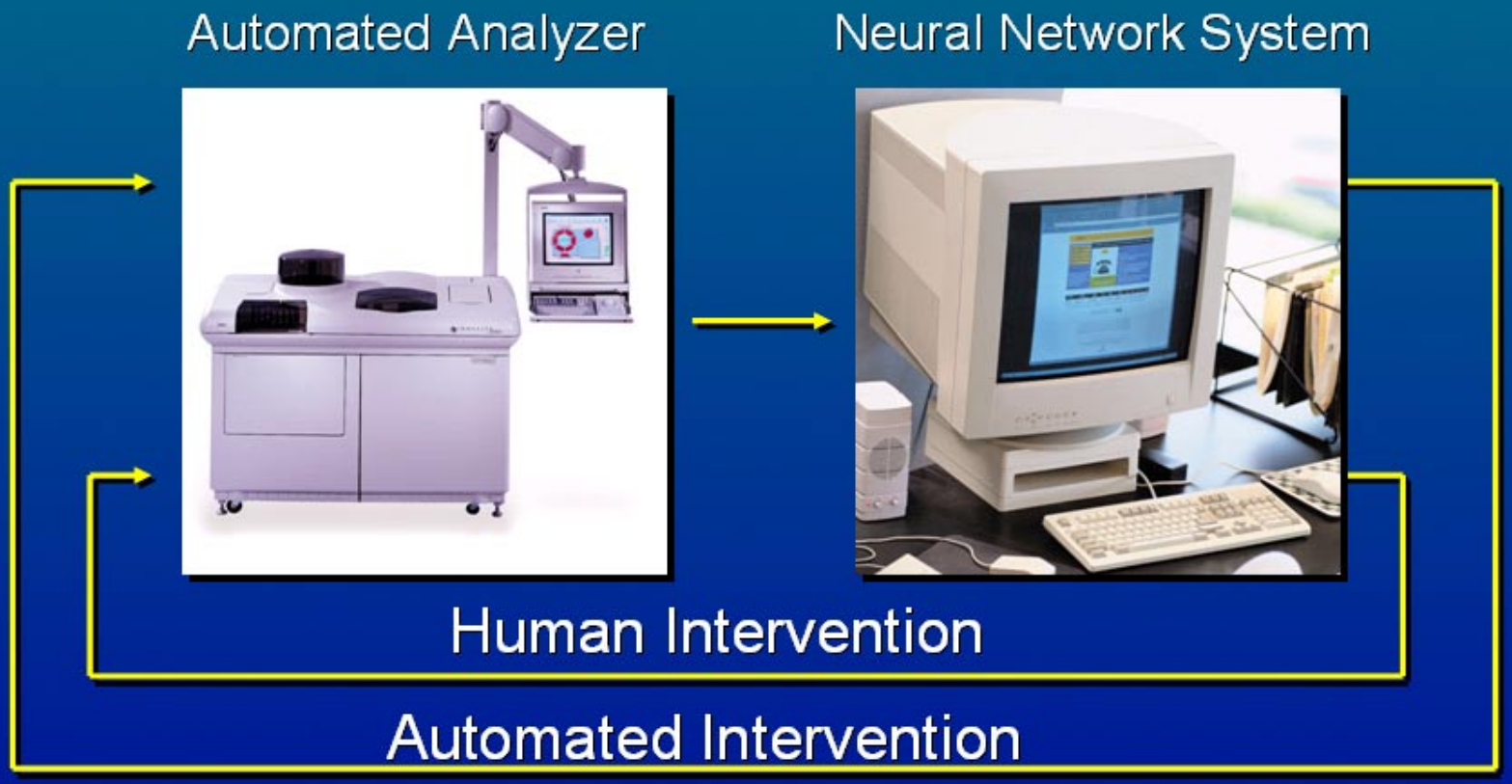
Abstract

We are developing a computational methodology that generates intelligent-agent ensembles for biomedical regression applications. In recall mode, such ensembles give a nonparametric distribution of answers rather than a single point estimate. These distributions may be useful for variance analysis and for discovering new domain knowledge, particularly when the estimated value is previously known. We illustrate our methodology with the problem of predicting serum analyte values with data from a clinical chemistry automated analyzer.

Introduction

A Chem-20 panel provides values for 20 analytes in a blood serum specimen. For a single specimen, we would like to predict each of the 20 analyte values from the accompanying 19 analyte values for the following reasons: (1) to detect spurious specimen and instrument problems, (2) for on-line monitoring of systematic lab testing problems, (3) for on-line monitoring of laboratory analyzer problems, (4) for predicting missing data for research studies, (5) to generalize what is learned to other kinds of applications. Instead of building one model for each held-out analyte, we built an ensemble of 1001 models to assess variance and to explore what new domain insights might be gained. We are using back error propagation as our modeling tool.

Neural Network System for On-line Laboratory Instrument QC Monitoring



Ensembling for Confidence

Adaptive learning methodologies such as artificial neural networks (ANNs) are usually applied to regression problems which subsume classification problems. "Ensembling" suggests many models instead of just one. There may be many reasons for ensembling. One reason is to estimate variance and confidence for predictive outcomes. One model gives only one point estimate, whereas an ensemble of say, 1001 models, gives a nonparametric distribution of answers from which confidence intervals may be derived.

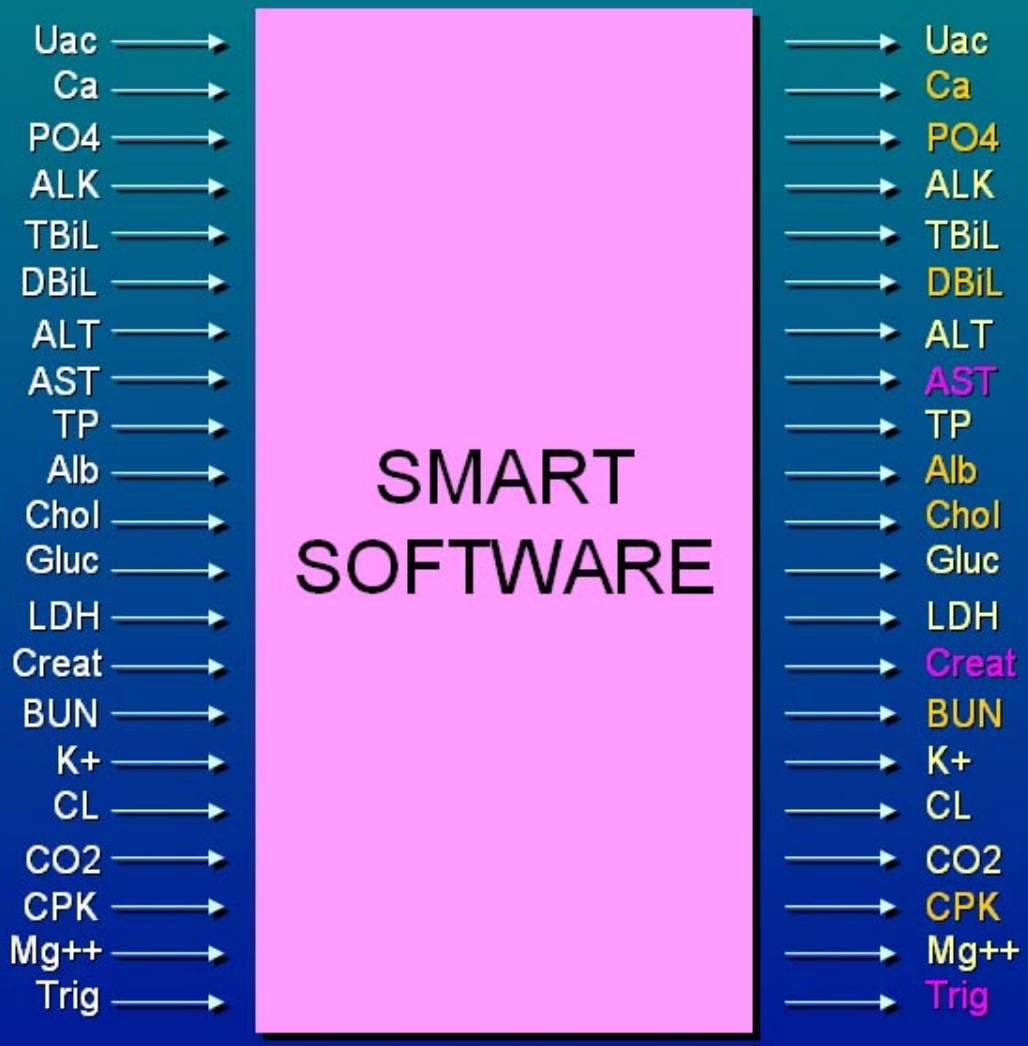
Basic Idea Regarding Ensembling

The basic idea of the work presented here as it relates to ensembling is as follows:

For a regression problem, develop not one but an ensemble, or cadre of same-kind models (intelligent-agents) as a way of assessing variance and discovering new knowledge. Each intelligent-agent in the ensemble, in general, will produce a slightly different result in recall mode because training is performed with random starting conditions (e.g. weights) and with random presentation of training cases (e.g. boot-strapping). In recall mode, when the result is known, the position of this result in the nonparametric distribution of the ensemble's predicted answers may reveal useful application domain knowledge.

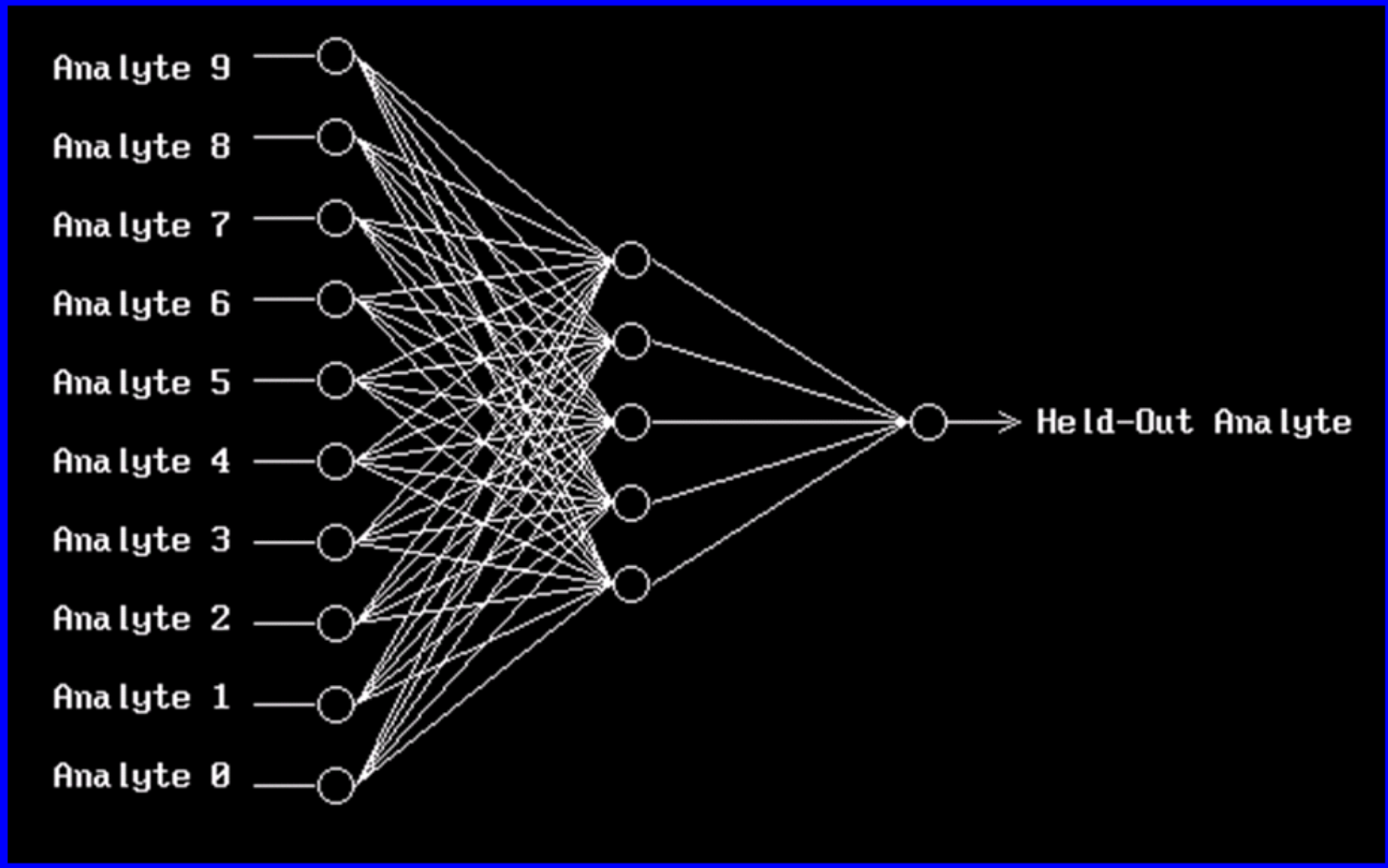
Training & Validating Ensembles

We trained and validated analyte-specific ANNs in which the value of the analyte of interest was held out for prediction with the remaining 19 analyte values. We used 400 cases to train and 74 cases to validate 20 analyte-specific ensembles of 1001 no hidden layer back error propagation ANNs. Data was scaled to the 0-1 interval by sigmoid transformation of the z-score. Each ANN was initialized with random weights and trained with bootstrap sampling for 1 cycle.


Uac
Ca
PO4
ALK
TBiL
DBiL
ALT
AST
TP
Alb
Chol
Gluc
LDH
Creat
BUN
K+
CL
CO2
CPK
Mg++
Trig
SMART
SOFTWARE
Uac
Ca
PO4
ALK
TBiL
DBiL
ALT
AST
TP
Alb
Chol
Gluc
LDH
Creat
BUN
K+
CL
CO2
CPK
Mg++
Trig

CHEM-20 Analytes

Albumin
Alk Phos
ALT/GPT
AST/GOT
Bilirubin, Direct
Bilirubin, Total
Calcium
CO_2, Total
Chloride
CK, Total
Creatinine
Glucose
LD
Magnesium
Phos, Inorganic
Potassium
Protein, Total
Sodium
BUN
Uric Acid


Analyte 9
Analyte 8
Analyte 7
Analyte 6
Analyte 5
Analyte 4
Analyte 3
Analyte 2
Analyte 1
Analyte 0
Held-Out Analyte

Data Scaling

Analyte values were scaled to the 0-1 interval by sigmoid transforming the z-score as follows:

$$z = (x - x_{mean}) / sd$$

$$x_s = 1 / [\, 1 + \exp(-z) \,]$$

To reverse-scale neural network output values back to analyte values:

$$z = - \log [\, (1 - x_s) / x_s \,]$$

$$x = x_{mean} + z \;\; sd$$

Ensemble Evaluation

The figure of merit (FOM) used to evaluate each ANN training and validation performance was the average absolute value of the differences between the scaled (0-1) computed and known values. The figures of merit used to evaluate each ensemble are the median and 90% confidence intervals for the FOMs associated with all ANNs in that ensemble. If computed and known values were random a FOM of .33 would result. Median FOMs for training ranged from .10 to .20 as shown in the following tables.

Training & Validation Ensemble Performances

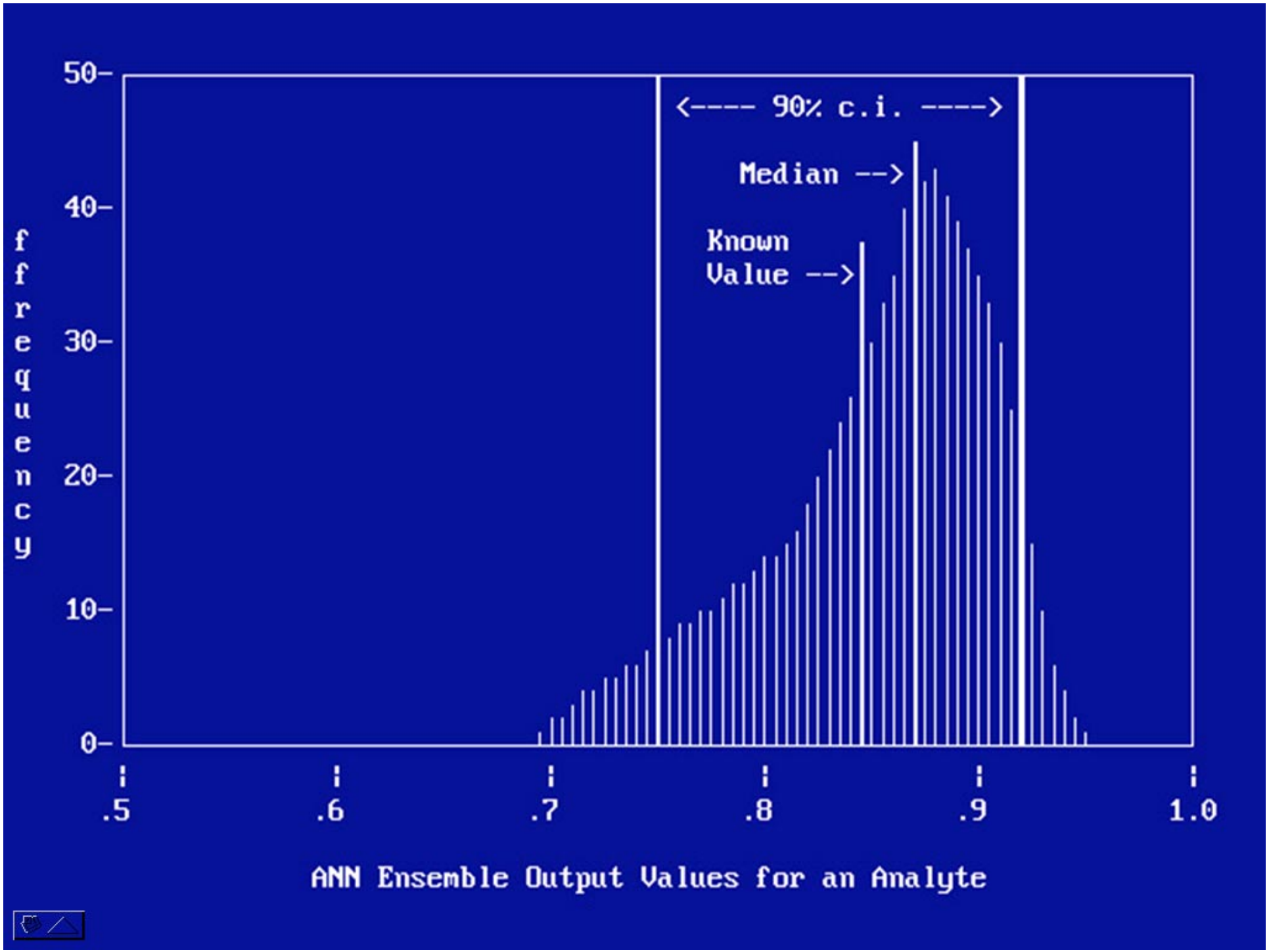
Analyte	Train	Valid	Analyte	Train	Valid
Albumin	.19	.__	Creatinine	.12	.__
Alk Phos	.16	.__	Glucose	.16	.__
ALT/GPT	.14	.__	LD	.14	.__
AST/GOT	.13	.__	Magnesium	.18	.__
Bilirubin, Direct	.14	.__	Phos, Inorganic	.18	.__
Bilirubin, Total	.17	.__	Potassium	.18	.__
Calcium	.18	.__	Protein, Total	.19	.__
CO_2, Total	.20	.__	Sodium	.19	.__
Chloride	.19	.__	BUN	.16	.__
CK, Total	.10	.__	Uric Acid	.20	.__

Top 10 Training Performers

Analyte	5%	50%	95%
CK,Total	.08	.10	.19
Creatinine	.09	.12	.19
AST/GOT	.10	.13	.21
Bilirubin, Direct	.11	.14	.21
LD	.11	.14	.21
ALT/GPT	.11	.14	.22
Glucose	.13	.16	.22
Protein, Total	.13	.16	.22
Alk, Phos	.14	.16	.24
Bilirubin, Total	.14	.17	.23

Bottom 10 Training Performers

Analyte	5%	50%	95%
Calcium	.14	.18	.24
Phos, Inorganic	.15	.18	.24
Potassium	.16	.18	.24
Magnesium	.16	.18	.25
Chloride	.16	.19	.24
Albumin	.15	.19	.25
Protein, Total	.15	.19	.25
Sodium	.16	.19	.25
CO_2 Total	.17	.20	.25
Uric Acid	.18	.20	.25


50-
40-
30-
20-
10-
0-
frequency
<---- 90% c.i. ---->
Median -->
Known
Value -->
.5
.6
.7
.8
.9
1.0
ANN Ensemble Output Values for an Analyte

CONCLUSIONS

We demonstrated work in progress toward developing a computational methodology that generates intelligent-agent ensembles for biomedical regression applications. We have illustrated this methodology with the problem of predicting serum analyte values with data from a clinical chemistry automated analyzer. The methodology has potential application in other areas of medicine as well as in other domains.

Jim DeLeo
Chief, Scientific Computing Section
Department of Clinical Research Informatics
Warren Grant Magnuson Clinical Center
National Institutes of Health
10 Center Drive
Building 10, Room 1C290
Bethesda, Maryland 20892-1172

Voice and Fax: 301-496-3848
E-mail: jdeleo@nih.gov
www.nih-bcig.org