Distributed Multivariate Regression Using Wavelet-based Collective Data Mining.

Daryl E. Hershberger and Hillol Kargupta

School of Electrical Engineering and Computer Science, Washington State University
Pullman, Washington 99164-2752, USA

E-mail: dhershbe and hillol @eecs.wsu.edu

This paper presents a method for distributed multivariate regression using wavelet-based Collective Data Mining (CDM). The method seamlessly blends machine learning and the theory of communication with the statistical methods employed in parametric multivariate regression to provide an effective data mining technique for use in a distributed data and computation environment. The technique is applied to two benchmark data sets, producing results that are consistent with those obtained by applying standard parametric regression techniques to centralized data sets. Evaluation of the method in terms of model accuracy as a function of appropriateness of the selected wavelet function, relative number of non-linear cross-terms, and sample size demonstrates that accurate parametric multivariate regression models can be generated from distributed, heterogeneous, data sets with minimal data communication overhead compared to that required to aggregate a distributed data set. Application of this method to Linear Discriminant Analysis, which is related to parametric multivariate regression, produced classification results on the Iris data set that are comparable to those obtained with centralized data analysis.

Key Words: data mining, distributed data mining, collective data mining, knowledge discovery, wavelets, regression

1. INTRODUCTION

This paper presents an approach to distributed multivariate regression using wavelet-based *Collective Data Mining* (CDM) [25]. CDM is an approach to Distributed Data Mining (DDM) that addresses difficulties introduced when distributed data sites observe heterogeneous sets of features.

Distributed data mining deals with methods of finding data patterns in a distributed data and computation environment. DDM methods allow distributed data to be analyzed with minimal data communication. Generally, DDM algorithms start with local data analysis followed by generation of a global model based on combining the results of the local analysis. In the general case, where different sites observe different sets of features, naive approaches to local analysis may be ambiguous and incorrect, resulting in incorrect global

models. CDM provides a well-grounded methodology to address this general case, offering an approach to the analysis of distributed, heterogeneous databases with distinct feature spaces.

The foundation of CDM is the observation that any function may be represented in distributed fashion by using an appropriate set of basis functions. Communication theory provides that efficient transmission of information is facilitated through the use of orthogonal functions [17]. Wavelet analysis techniques [37] provide a powerful tool for generating orthogonal basis function sets for use in CDM.

Parametric Multivariate Regression (MR) is a widely used statistical data analysis technique that can also be viewed as a supervised learning algorithm. The distributed MR technique presented here learns local information in terms of the coefficients of an orthogonal basis function representation, transmits a small (relative to the sample size) number of significant coefficients to a central site, and then generates a global model directly from that small set of significant coefficients. The method seamlessly blends machine learning and the theory of communication with the statistical methods employed in MR to provide an effective data mining technique for use in a distributed data and computation environment.

Section 2 begins with a description of the general DDM problem of heterogeneous data sets. This is followed by a review of related DDM work and an overview of MR. The section concludes with an example of the specific problem of naive data analysis within the context of parametric regression models in a DDM environment. Section 3 provides an overview of the foundations of CDM and a description of the wavelet techniques used for the distributed MR model. An algorithm for distributed MR using wavelet-based CDM is presented in Section 4. The performance of this CDM-MR method is then characterized using real "benchmark" data sets and larger synthetic data sets. Section 5 describes the application of the CDM regression model to Linear Discriminant Analysis (LDA) which is related to parametric multivariate regression. Section 6 summarizes the CDM work presented here for MR models and LDA, and discusses future research directions.

2. BACKGROUND

This section presents background material related to MR models using wavelet-based CDM. A simple model of distributed, heterogeneous, data sites is explained first. This is followed by a review of related DDM research. Next a brief overview of MR is provided and finally an example that demonstrates the incorrect results that may be obtained by naive application of parametric regression techniques to distributed data is presented.

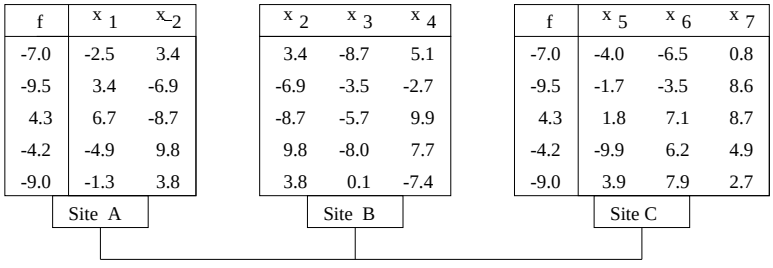

FIG. 1. Distributed data sites with a vertically partitioned feature space.

2.1. Motivation

DDM deals with the problem of finding data patterns in an environment with distributed data and computation. A typical application domain of DDM either has inherently distributed data sources or centralized data partitioned at different sites. The data sites may be homogeneous, i.e. each site stores data for exactly the same set of features. In the general case, however, the data sites may be heterogeneous, each site maintaining databases with different kinds of information. For example, financial institutions (banks, insurance companies, credit-card issuers) wish to combat fraud and posses data that would be of use to one another in this effort. However, customer privacy concerns prevent this data from being combined into a single data base. CDM allows pertinent information and data patterns to be extracted from the individual data bases without compromising customer privacy. Another example of distributed heterogeneous data sets is the local data associated with a large set of distributed sensors (either environmental or industrial) where the sensed parameters are different. There may be no organizational barriers to centralizing this data but time and processing constraints may give CDM an advantage over centralized techniques.

In the general case the feature sets observed at different sites are different. A simple example of this is the case of a *vertically partitioned data set*. Figure 1 illustrates this situation. While this paper considers the problem of developing MR models for this simple case of vertically partitioned data sets, CDM is applicable to the general case of heterogenous data sets and it is expected that CDM-MR may be extended to this general case.

Given a set of observations representing discretely sampled continuous valued features, the task is to use parametric regression techniques to learn a function that estimates the unknown value of a dependent feature as a function of other observed independent features. The given set of observed feature values is sometimes called the training data set. In Figure 1 the column for f denotes the feature value to be estimated; $x_1, x_2, x_3, x_4, x_5, x_6$ and x_7 denote the independent features that are used to estimate f. The data sets available at the different sites are used as the data that the regression is performed on. If the f column is not observed everywhere and it is required to learn the local models it is broadcasted to every site.

One requirement for implementation of DDM algorithms with vertically partitioned data sets is a method for properly aligning the feature pattern vectors in different partitions with each other. This requirement introduces a certain amount of flexibility in defining relations among feature sets. One possibility is to use an operation similar to the Join operation of relational data bases. Note that in Figure 1 site A and B share feature x_2 while site A and C share feature f. The alignment of the feature pattern vectors at sites A and B could be accomplished by a Join based on x_2 followed by alignment with the pattern vectors at site C using Join based on f. If no common feature exists then an association must be made based on some prior knowledge or expectation regarding the model under development.

As Figure 1 shows each site may observe features such that a majority are unique to that site and therefore the sites are called heterogeneous. There exists little work for this general case of DDM. The following section reviews related work in DDM.

2.2. Related Work

This section briefly reviews some of the existing DDM work. This work may be grouped into four basic categories, Meta-learning and Stacking, Collective Data Mining,

Distributed Association Rule Learning, and other DDM techniques. Following these several experimental DDM systems are reviewed.

Meta-learning [6, 5, 7] and Stacking [38] are examples of techniques for mining homogeneous distributed data. In the meta-learning approach supervised learning techniques are first used to detect concepts at local data sites, then meta-level concepts are learned from a data set generated using the locally learned concepts, resulting in a meta-classifier. Different inductive learning algorithms may be employed to learn the local concepts, and the meta-level learning may be applied recursively, producing a hierarchy of meta-classifiers. The JAM system [33] is a meta-learning base distributed data mining framework that has been used for fraud detection in the banking domain [28].

Collective data mining [24, 25] address the issues associated with mining heterogeneous data sites. At the foundation of CDM is the observation that any function may be represented in a distributed manner using an appropriate set of basis functions. By using orthogonal basis function, correct models of local information may be developed in terms of the basis function coefficients. A global model may be generated by communicating a small fraction of the local basis coefficients to a central site. Learning algorithms that have been applied to CDM include decision trees and the parametric multivariate regression techniques presented in this paper.

The mining of association rules in distributed data bases has been examined in [9]. In this work the *Distributed Mining of Association rules* (DMA) algorithm is presented. This algorithm takes advantage of the inherent parallel environment of a distributed database as opposed to previous works that tended to be sequential in nature.

The *fragmented approach* to mining classifiers from distributed data sources is suggested by [10]. In this method a single, best, rule is generated in each distributed data source. These rules are then ranked using some criterion and some number of the top ranked rules are selected to form the rule set. In [27] the authors extend efforts to automatically produce a Bayesian belief network from discovered knowledge by developing a distributed approach to this exponential time problem.

In [39] the author presents two models of distributed Bayesian learning. Both models employ distributed agent learners each of which observes a sequence of examples and produces an estimate of the parameter specifying the target distribution and a population learner that combines the output of the agent learners in order to produce a significantly better estimate of the parameter of the target distribution. One model applies to situations in which the agent learners observe data sequences generated according to the identical target distribution while the second model applies when the data sequences may not have the identical target distribution over all agent learners.

The PADMA system [23, 22] achieves scalability by locating agents with the distributed data sources. An agent coordinating facilitator gives user requests to local agents that then access and analyze local data, returning analysis results to the facilitator which merges the results. The high level results returned by the local agents are much smaller than the original data thus allowing economical communication and enhancing scalability. The authors report on a PADMA implementation for unstructured text mining but note that the architecture is not domain specific.

Papyrus, a system in development by the National Center for Data Mining [16], is a hierarchical organization of the nodes within a data mining framework. The intent of this project is to develop a distributed data mining system that reflects the current distribution

of the data across multiple sites and the existing network configurations connecting these configurations.

Another system under development [36] concerns itself with the efficient decomposition of the problem in a distributed manner and utilizes clustering and Expected Maximization algorithms for knowledge extraction.

Work has also been done concerning using the Internet [8] as the framework for large scale data mining operations. This work is also applicable to intra-nets, and addresses issues of heterogeneous platforms and security issues.

The WoRLD system [1] for inductive rule-learning from multiple distributed databases uses *spreading activation* instead of item-by-item matching as the basic operation of the inductive engine. Database items are labeled with markers (indicating in or out of concept) that are then propagated through databases looking for values where in or out of concept markers accumulate.

A basic requirement of algorithms employed in DDM is that they have the ability to scale up. A survey of methods of scaling up inductive learning algorithms is presented in [32].

2.3. Overview of Multivariate Regression

MR is a widely used data analysis technique owing to its ease of use and intuitive theoretical basis [30, 13]. MR involves fitting a parametric function model to a set of data. In this sense it is a form of inductive supervised learning.

The functions analyzed have the form $f = b_1 r_1 + b_2 r_2 + \ldots + b_k r_k$ where the b_i-s are constant coefficients and the r_i terms are linear or non-linear functions of the feature set. For example, if the feature set contains features x_a and x_b, then x_a, x_b^2, and $\log(\frac{x_b}{x_a})$ may be present in the r_i terms of the function.

Given a data set Ω consisting of samples values of the features in the feature set and the associated function value, possibly containing some random error, the objective of MR is to produce an estimate $\hat{f} = \hat{b}_1 r_1 + \hat{b}_2 r_2 + \ldots + \hat{b}_k r_k$ of the function f where the regression model coefficients $\hat{b}_i$ are estimates of the b_i. The technique used in MR is to find the set of $\hat{b}_i$-s that minimize $\sum_\Omega (f - \hat{f})^2$, the sum of the squares of the difference between the sample and estimated function value over the data sample set.

Using matrix notation to represent the function relation in terms of a data set of size N gives

$$\mathbf{F} = \mathbf{XB} + \epsilon$$

where $\mathbf{F}$ is a $N \times 1$ matrix of function sample values, $\mathbf{X}$ is a $N \times k$ matrix where each column represents the sample data for one independent feature or regressor and each row contains the set of observed values of the independent features for one sample, $\mathbf{B}$ is a $k \times 1$ matrix of the b_i-s, and ϵ is a $N \times 1$ matrix of values representing errors in the measured value of f. If the matrix $\mathbf{X^T X}$ is invertible then the minimum squared error estimate of the regression coefficients is

$$\hat{\mathbf{B}} = (\mathbf{X^T X})^{-1} \mathbf{X^T F}$$

where $\hat{\mathbf{B}}$ is a $k \times 1$ matrix of $\hat{b}_i$-s.

The following section provides an example of the problems that may arise from naive application of MR to vertically partitioned data sets.

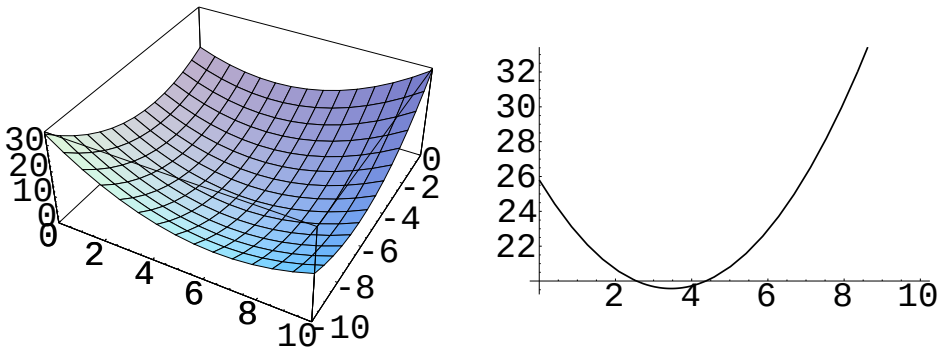

FIG. 2. (Left) The global error function. (Right) The local error function at site A.

2.4. The Naive Approach to Local Regression Models

Data modeling is a mature field that has many well-understood techniques in its arsenal, including parametric regression. However, like many of these traditional techniques, parametric regression cannot be directly used in a distributed environment with a vertically partitioned feature space. The following example demonstrates that even a simple decomposable parametric regression problem with no measurement error can produce misleading results in a distributed environment. Consider the function, $g(x_1, x_2) = 5x_1 - 5x_2$, where x_1 and x_2 are real valued variables, and the sample data set $\Omega = \{(1.0, 1.0, 0.0), (-0.2, -0.2, 0.0), (0.2, -1.0, 6.0), (-1.0, 0.2, -6.0)\}$, where each entry is of the form $(x_1, x_2, g(x_1, x_2))$. Let us try to fit a model, $\hat{g}(x_1, x_2) = \hat{b}_1 x_1 + \hat{b}_2 x_2$, to this data by minimizing the mean-square error. The overall mean square error computed over the data set Ω is, $\frac{1}{4}\sum_{x_1,x_2\in\Omega}(g-\hat{g})^2 = 0.52(5-\hat{b}_1)^2+0.52(-5-\hat{b}_2)^2+0.32(5-\hat{b}_1)(-5-\hat{b}_2)$ Figure 2 (Left) shows the error surface with a global minima at $\hat{b}_1 = 5$ and $\hat{b}_2 = -5$. It is a simple quadratic function and finding the minima is quite straight forward.

Now let us consider the data set to be vertically partitioned; meaning x_1 is observed at site A and x_2 is observed at a different site, B. Let us choose a linear model $\hat{g}(x_1) = \hat{b}_1 x_1$. The mean square error function for site A is $\frac{1}{4}\sum_{x_1,x_2\in\Omega}(f - \hat{f})^2 = 0.52(5 - \hat{b}_1)^2 + 20.8 - 1.6(5 - \hat{b}_1)$. Figure 2 (Right) shows this local error function. It clearly shows that the minima of this error function is not the same as the globally optimal value of $\hat{b}_1$, i.e. 5. This example demonstrates that even for simple linear data and models naive approaches to minimize mean-square error may be misleading in a distributed environment. As will be demonstrated later in this paper, wavelet-based CDM offers a correct, viable solution to this problem.

3. COLLECTIVE DATA MINING AND WAVELET BASIS

This section provides an overview of the foundations of CDM and the wavelet basis used for the distributed MR model developed later.

3.1. The Foundations of Collective Data Mining

A thorough treatment of the foundations of CDM is presented in Kargupta et. al. [25]. What follows in this section is an abbreviated summary of that work as it applies to distributed MR.

In MR relations among the different members of the domain and the corresponding members in the range (class labels or the output function values, denoted by Y) are desired. The goal is to learn a function $\hat{f} : X^L \rightarrow Y$ from the data set $\Omega = \{(\bar{x}_{(1)}, y_{(1)}), (\bar{x}_{(2)}, y_{(2)}), \cdots (\bar{x}_{(N)}, y_{(N)})\}$ generated by an underlying function $f : X^L \rightarrow Y$ such that the $\hat{f}$ approximates f. Individual members of the domain $\bar{x} = x_1, x_2, \cdots, x_L$ are L-tuples and the x_ℓ-s correspond to individual features of the domain.

The foundation of CDM is based on the fact that by using an appropriate set of basis functions any function can be represented in a distributed fashion. Let Ξ be a possibly infinite set of basis functions. Associate an index with each basis function in Ξ and denote the k-th basis function in Ξ by Ψ_k and the set of all such indices of the basis functions as Ξ_I. The function $f(\bar{x})$ can be represented as

$$f(\bar{x}) = \sum_{k \in \Xi_I} w_k \Psi_k(\bar{x}) \tag{1}$$

where w_k denotes the coefficient of the k-th basis function. The objective is to generate a function $\hat{f}(\bar{x})$ that approximates $f(\bar{x})$ from a given data set

$$\hat{f}(\bar{x}) = \sum_{k \in \hat{\Xi}_I} \hat{w}_k \Psi_k(\bar{x}) \tag{2}$$

where $\hat{\Xi}_I$ denotes a subset of Ξ_I and $\hat{w}_k$ denotes the approximate estimation of the coefficients w_k. For distributed MR using wavelet-based CDM the underlying task is essentially to compute the significant wavelet coefficients, $\hat{w}_k$-s.

If a function has a large number of significant basis coefficients exponential time (in number of features) is required for computing the orthonormal representation. In order to have polynomial time computation of the coefficients, two conditions must be met: 1) a sparse representation where most of the coefficients are zero or negligible, and 2) approximate evaluation of the significant coefficients.

In most MR models non-linearity typically remains bounded so not all the features non-linearly interact with every other feature. It is normally acceptable to assume that the number of features that non-linearly interact with any given feature is bounded by some constant. If this is not true then the problem is completely non-linear and is likely to be difficult for even a centralized data mining algorithm let alone DDM. This requirement has a deep root in issues of polynomial time, probabilistic and approximate learn-ability [26]. Bounded non-linearity assures that the orthonormal representation will be sparse, satisfying the first condition of polynomial time computation.

The second condition is associated with the fact that only a sample from the domain will be available for computing the basis coefficients. This will not cause a problem as long as our sample size is reasonable. To illustrate the rationale behind this observation consider what happens when both sides of Equation 1 are multiplied by $\Psi_j(\bar{x})$ resulting in $f(\bar{x})\Psi_j(\bar{x}) = \sum_{k \in \Xi_I} \omega_k \Psi_k(\bar{x})\Psi_j(\bar{x})$. If the sample data set is denoted by Ω, then by

summing both side over all members of Ω the result is

$$\sum_{\bar{x}\in\Omega} f(\bar{x})\Psi_j(\bar{x}) = \sum_{\bar{x}\in\Omega}\sum_{k\in\Xi_I} \omega_k \Psi_k(\bar{x})\Psi_j(\bar{x}) \tag{3}$$

Since $\Psi_j(\bar{x})\Psi_j(\bar{x}) = 1$ then $\sum_{\bar{x}\in\Omega} \Psi_j(\bar{x})\Psi_j(\bar{x}) = |\Omega|$ where $|\Omega|$ is the sample size and it follows that

$$\frac{1}{|\Omega|}\sum_{\bar{x}\in\Omega} f(\bar{x})\Psi_j(\bar{x}) = \omega_j + \sum_{k\in\Xi_I, k\neq j} \omega_k \frac{\sum_{\bar{x}\in\Omega} \Psi_k(\bar{x})\Psi_j(\bar{x})}{|\Omega|}$$

If the population mean over the complete domain is zero then the sample mean must approach zero as the sample size increases. Since $\frac{1}{|\Omega|}\sum_{\bar{x}\in\Omega} \Psi_k(\bar{x})\Psi_j(\bar{x})$ is the sample mean it follows that for large sample sizes (typically the case for data mining problems) the last term should approach zero.

In summary the primary steps of the CDM algorithm are [21]

1. generate appropriate orthonormal basis coefficients at each local site;

2. if needed, move an appropriately chosen sample of the data sets from each site to a single site and generate the approximate basis coefficients corresponding to non-linear cross terms;

3. combine the local models, transform the model into the user described canonical representation, and output the model.

It should be noted that the approach to distributed MR using wavelet-based CDM does not require the transfer of raw data (feature sample values) to a central site in order to estimate non-linear cross terms. These estimates are instead generated directly from the local model orthonormal basis coefficients.

The following section introduces the wavelet methods [37] used to create the orthonormal basis representation needed for implementing distributed MR in CDM.

3.2. Wavelet Basis and Wavelet-Packet Analysis

A wavelet basis consists of a set of scaling basis functions ϕ_k and a set of wavelet basis functions ψ_k. The wavelet functions are dilated and translated versions of the scaling functions [34, 35, 37]. The relation between the scaling and wavelet functions may be understood by considering a vector space S^j with 2^j dimensions defined on the interval $[0, 1)$. Note that S^j contains all functions that are piece-wise constant on 2^j equal sub-intervals defined on the interval $[0, 1)$. Since S^{j-1} is also defined on $[0, 1)$ every function in S^{j-1} is also in S^j with each interval in S^{j-1} considered to correspond to two contiguous intervals in S^j. Let $S^j = S^{j-1} + W^{j-1}$ where the subspace W^{j-1} is the orthogonal complement of S^{j-1} in S^j. If the basis functions for S^{j-1} are the scaling functions ϕ_k^{j-1} then the basis functions for W^{j-1} will be the wavelet functions ψ_k^{j-1}. Since S^{j-1} and W^{j-1} are complementary orthogonal spaces the ϕ_k^{j-1} and ψ_k^{j-1} will be orthogonal to each other in S^j. Now if the ϕ_k^{j-1} form an orthogonal basis for S^{j-1} and the ψ_k^{j-1} form an orthogonal basis for W^{j-1} then combined the ϕ_k^{j-1} and ψ_k^{j-1} will form an orthogonal basis for S^j.

This work employs a simple set of scaling functions for S^j, the scaled and translated "box" functions [34] defined on the interval $[0, 1)$ by

$$\phi_k^j(x) = \phi(2^j x - i), k = 0, \ldots, 2^j - 1,$$

where

$$\phi(x) = \begin{cases} 1 & \text{for } 0 \leq x < 1 \\ 0 & \text{otherwise} \end{cases}$$

The wavelet functions corresponding to the box basis functions are the *Haar wavelets*

$$\psi_k^j(x) = \psi(2^j x - i), i = 0, \ldots, 2^j - 1,$$

where

$$\psi(x) = \begin{cases} 1 & \text{for } 0 \leq x < \frac{1}{2} \\ -1 & \text{for } \frac{1}{2} \leq x < 1 \\ 0 & \text{otherwise} \end{cases}$$

A function in S^j may be represented in terms of these basis functions as

$$f(x) = s_0^j \phi_0^j + s_1^j \phi_1^j + \ldots + s_{2^j-1}^j \phi_{2^j-1}^j$$

or also as,

$$f(x) = s_0^{j-1} \phi_0^{j-1} + \ldots + s_{2^{j-1}-1}^{j-1} \phi_{2^{j-1}-1}^{j-1} + d_0^{j-1} \psi_0^{j-1} + \ldots + d_{2^{j-1}-1}^{j-1} \psi_{2^{j-1}-1}^{j-1}$$

The wavelet coefficients, s_k^{j-1} and d_k^{j-1} are generated by convolution of the s_k^j with a set of orthogonal quadrature filters, H and G. For the Haar wavelets, $H = \{\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\}$ and $G = \{\frac{1}{\sqrt{2}}, \frac{-1}{\sqrt{2}}\}$.

The *Wavelet-Packet* transform of a function in S^j is calculated by recursively applying the quadrature filters to the s and d coefficients of the next lower scale space and wavelet space as if each represented a separate scale space. Figure 3 shows how the wavelet-packet transform is generated by recursively applying the quadrature filters to the scale and wavelet subspaces. The vector $\bar{S}^j = s_0^j, s_1^j, \ldots, s_{2^j-1}^j$ input at the top level of the wavelet-packet decomposition is formed by assuming that the observed values of the piece-wise constant function represent the coefficients of the scaling functions in S^j. If the original function is in S^j then 2^j orthogonal subspaces $S_k^0, k = 0, \ldots, 2^j - 1$, will result from j recursive applications of H and G.

Using the Haar function wavelets partitions S^j into 2^j orthogonal subspaces S_k^0 that are each spanned by one of the first 2^j Walsh functions [17]. Therefore, the Walsh functions form an orthogonal basis, $\Psi_0, \Psi_1, \ldots, \Psi_{2^j-1}$, for S^j and the Walsh coefficients, $\omega_k : k = 0, 2^j - 1$ are equivalent to the wavelet-packet coefficients of the 2^j orthogonal subspaces S_k^0.

3.3. The Wavelet-based CDM Approach to Local Regression

To demonstrate the wavelet-based CDM regression method consider again the naive example given in Section 2.3. Table 1 shows the Walsh coefficients, ω_k obtained from the wavelet-packet decomposition of x_1, x_2, and $g(x_1, x_2)$. Again selecting regression models $\hat{g}(x_1) = \hat{b}_1 x_1$, $\hat{g}(x_2) = \hat{b}_2 x_2$ and applying standard MR techniques to the non-zero Walsh coefficient sets in each partition gives,

$$\hat{b}_1 = \omega_2^{(g)} / \omega_2^{(x_1)} = 3.0/0.6 = 5.0$$

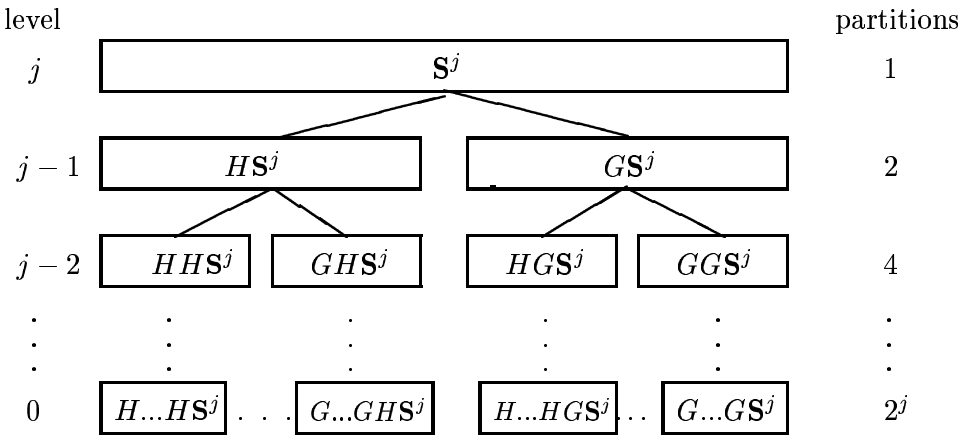

FIG. 3. Application of quadrature filters in wavelet-packet decomposition.

and

$$\hat{b}_2 = \omega_3^{(g)}/\omega_3^{(x_2)} = -3.0/0.6 = -5.0$$

Note that 1) the representation of the information embodied in the example has become sparse compared to the original, 2) the regression coefficients for this decomposable problem may now be determined directly from the non-zero wavelet coefficients in each partition without recourse to information exchange.

In the general case some information in the form of wavelet coefficients will need to be communicated among partitions in order to resolve non-linear terms.

TABLE 1

Walsh (wavelet-packet) coefficients for naive example.

k	$\omega_k^{(g)}$	$\omega_k^{(x_1)}$	$\omega_k^{(x_2)}$
0	0.0	0.0	0.0
1	0.0	0.4	0.4
2	3.0	0.6	0.0
3	-3.0	0.0	-0.6

4. MULTIVARIATE LINER REGRESSION

Parametric regression [30], is a form of supervised learning that is applicable to CDM. In this section one approach to distributed MR based on an orthogonal wavelet basis is presented. We begin with a description of the method used to generate local models, followed by the method for generating the global model. Next we apply these methods to two benchmark data sets and compare the resulting model statistics with those obtained using standard MR techniques on centralized data sets. Following this, large synthetic data sets are employed to characterize model performance and scalability. Finally, we address the overall performance bounds for the methodology.

4.1. Generating Local Models

One of the keys to CDM is that the local models represent local information in terms of the coefficients of a function set that forms an orthogonal basis for the distributed data set. In the case of distributed MR the coefficients of interest are the Walsh (wavelet-packet) coefficients obtained by performing a wavelet-packet decomposition on the samples of each feature.

Given a partitioned set of real-valued features $\bar{x} = x_1, x_2, \ldots, x_\ell$ and a data set of these features with N samples $\Omega = \bar{x}_{(1)}, \bar{x}_{(2)}, \ldots, \bar{x}_{(N)}$ let $\bar{x}^A$ be the set of features and Ω_A be the associated data set found in partition A. If $x_1 \in \bar{x}^A$ then let $\bar{x}^{(1)}$ denote the column vector of dimension N formed from the sample values of feature x_1 in partition A and $x_1(n)$ be the n^{th} element of $\bar{x}^{(1)}$.

Each of the $\bar{x}^{(i)} \in \bar{x}_A$ may be transformed into a set of wavelet basis coefficients $\omega_k^{(x_i)}$ using wavelet-packet decomposition. In terms of the wavelet-packet transform technique,

$$x_i(n) = x_i(0)\phi_0^j + x_i(1)\phi_1^j + \ldots + x_i(N)\phi_N^j$$

therefore $\bar{x}^{(i)} = \bar{S}^j$ where in this case $N = 2^j$. Once the wavelet-packet transform of $\bar{x}^{(i)}$ is performed $x_i(n)$ may be expressed in terms of the Walsh basis as

$$x_i(n) = \sum_{k=1}^{N} \omega_k^{(x_i)} \Psi_k(n) \tag{4}$$

If the wavelet functions are properly selected based on the feature sample characteristics the representation of the feature in the wavelet basis will be sparse and many of the ω_k-s will be zero or insignificant. Note that the feature samples, $x_i(n)$, may be exactly re-constituted from the wavelet coefficients using Equation 4. The wavelet representation may be made more sparse by zeroing coefficients that have an absolute value below some threshold or alternatively retaining only a fixed number or percentage of the coefficients with the largest absolute values. If the Walsh coefficients are ordered from largest to smallest absolute value and then the first $M < N$ coefficients are retained Equation 4 becomes,

$$\hat{x}_i(n) = \sum_{j=1}^{M} \omega_j^{(x_i)} \Psi_j(n)$$

where the index j is on the ordered coefficients. This *thresholding* process will eliminate the ability to exactly re-constitute the feature sample values from the wavelet coefficients. However, a feature sample set estimate re-constituted from a set of thresholded coefficients will be the minimum square error (MSE) estimate of the original feature samples based on the number of wavelet coefficients retained [3]. The wavelet coefficients that remain after thresholding form the local model.

4.2. Generating a Global Multivariate Regression Model

The local model coefficients contain all of the information needed from the sample data sets in order to generate a global model. These local model coefficients, representing a MSE estimate of the information available in the original feature samples, are communicated to a central site to facilitate the global model generation process. In order to generate the global model the wavelet coefficients for cross-terms and higher-order terms, the r_i terms

in the function to be fit, must first be estimated. This may be accomplished directly from the local model wavelet coefficients.

To see that the wavelet coefficients for the cross-terms and higher-order terms may be calculated directly from the local model wavelet coefficients first recall that the wavelet basis functions are an orthogonal function set so that

$$\frac{1}{N}\sum_{n=0}^{N-1}\Psi_i(n)\Psi_j(n) = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

where the $\Psi(n)$ are particular basis functions and N is the sample column vector space size. Certain sets of orthogonal basis functions such as Walsh are closed under product such that

$$\Psi_i(n)\Psi_j(n) = \Psi_k(n) \tag{6}$$

For the case of the Walsh basis any set of Walsh functions satisfying equation 6 are related by

$$k = i \oplus j \tag{7}$$

where $\oplus$ represents addition modulo 2 of the binary representations of the index values i and j. Therefore in Equation 6 $k = 0$ only if $i = j$, so it follows that

$$\frac{1}{N}\sum_{n=0}^{N-1}\Psi_i(n) = \begin{cases} 1 & \text{if } i = 0 \\ 0 & \text{otherwise} \end{cases}$$

For feature x_a the coefficient $\omega_k^{(x_a)}$ for the k^{th} Walsh basis function is given by

$$\omega_k^{(x_a)} = \frac{1}{N}\sum_{n=0}^{N-1} x_a(n)\Psi_k(n)$$

and given a complete set of basis coefficients an individual sample value of feature x_a may be calculated using

$$x_a(n) = \sum_{k=0}^{N-1}\omega_k^{(x_a)}\Psi_k(n) \tag{8}$$

For the cross-term $r_{ab} = x_a x_b$ the relation for calculating the coefficient of the k^{th} basis function is

$$\omega_k^{(r_{ab})} = \frac{1}{N}\sum_{n=0}^{N-1} x_a(n)x_b(n)\Psi_k(n)$$

and by substituting the relation for x_a and x_b based on Equation 8

$$\omega_k^{(r_{ab})} = \frac{1}{N}\sum_{n=0}^{N-1}\left(\sum_{i=0}^{N-1}\omega_i^{(x_a)}\Psi_i(n)\right)\left(\sum_{j=0}^{N-1}\omega_j^{(x_b)}\Psi_j(n)\right)\Psi_k(n)$$

Rearranging terms gives

$$\omega_k^{(r_{ab})} = \frac{1}{N} \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} \omega_i^{(x_a)} \omega_j^{(x_b)} \sum_{n=0}^{N-1} \Psi_i(n) \Psi_j(n) \Psi_k(n)$$

Now note that the last sum in this relation will be zero unless $\Psi_i \Psi_j = \Psi_k$ or, equivalently, $i \oplus j = k$. The sum may be replaced by the Dirac δ-function $\delta(x)$

$$\delta(x) = \begin{cases} 1 & \text{if } x = 0 \\ 0 & \text{otherwise} \end{cases}$$

The relation for the cross-term basis function coefficient becomes

$$\omega_k^{(r_{ab})} = \frac{1}{N} \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} \omega_i^{(x_a)} \omega_j^{(x_b)} \delta((i \oplus j) - k)$$

where again $\omega_i^{(x_a)}$ and $\omega_j^{(x_b)}$ are wavelet-basis coefficients for features x_a and x_b, $\delta((i \oplus j) - k)$ is the Dirac delta function and N is the vector space dimension (the number of samples). The relation for the cross-term basis coefficient is now in terms of only the basis coefficients of the cross-term features.

Once the global model wavelet coefficients are available the regression coefficients may be estimated by performing a regression directly on the wavelet coefficients. This is possible since the wavelet-packet transform is a linear transform. Thus the original function estimation relation given in Section 2.3

$$\hat{f} = \hat{b}_1 r_1 + \hat{b}_2 r_2 + \ldots + \hat{b}_k r_k$$

is transformed linearly to

$$\hat{\omega}_j^{(f)} = \hat{b}_1 \omega_j^{(r_1)} + \hat{b}_2 \omega_j^{(r_2)} + \ldots + \hat{b}_N \omega_j^{(r_k)}$$

where $\omega_j^{(r_i)}$ is the coefficient of the j^{th} basis function in the orthogonal representation of the term r_i of the function to be fit and the $\hat{b}_i$ coefficients we wish to estimate are the same in both relations. Standard centralized MR techniques [30, 13] are used on the set of wavelet coefficients $\omega_j^{(r_i)}$ to find the set of $\hat{b}_i$-s that minimize $\sum_M \left(\omega_j^{(f)} - \hat{\omega}_j^{(\hat{f})}\right)^2$.

In Section 2 we noted that in terms of matrix notation, the estimates of the regression coefficients could be calculated from the sample data as

$$\hat{\mathbf{B}} = (\mathbf{X^T X})^{-1} \mathbf{X^T Y}$$

In [4] the authors provide a detailed description of numerical methods for manipulating matrices to solve this specific equation. For the purposes of the work presented here however a somewhat similar approach, that is also presented in [4] was employed. Gaussian elimination was used to solve the k simultaneous equations

$$\begin{array}{ccccccccc} c_{11}\hat{b}_1 & + & c_{12}\hat{b}_2 & + & \cdots & + & c_{1k}\hat{b}_k & = & c_{1f} \\ c_{21}\hat{b}_1 & + & c_{22}\hat{b}_2 & + & \cdots & + & c_{2k}\hat{b}_k & = & c_{2f} \\ \vdots & & \vdots & & \vdots & & \vdots & & \vdots \\ c_{k1}\hat{b}_1 & + & c_{k2}\hat{b}_2 & + & \cdots & + & c_{kk}\hat{b}_k & = & c_{kf} \end{array}$$

where

$$c_{ij} = \sum \omega_k^{r_i} \omega_k^{r_j} - \frac{\sum \omega_k^{r_i} \sum \omega_k^{r_j}}{N}$$

$$c_{if} = \sum \omega_k^{r_i} \omega_k^{f} - \frac{\sum \omega_k^{r_i} \sum \omega_k^{f}}{N}$$

It should be noted that while multivariate polynomial regression has been the model for the mathematical development presented here, in general the terms r_i of the function to be fit may be non-polynomial functions of the feature set. In these situations additional operations in the form of these functions must be applied to the observed feature samples or their wavelet coefficients. The most efficient order of application of the functions and wavelet transforms will depend on the form of the terms and the partitioning of the feature set.

As will be seen in the following section the accuracy of the global model decreases as the number of cross-terms and higher-order terms increases for a given number of features and level of information communicated. This performance issue may be offset by a modification to the method used to generate the local models. In partition A the terms $R_A(\bar{x}_k), k \in \Xi_A$ of the polynomial dependent only on features in partition A may be formed for each sample and the wavelet-packet decomposition calculated for each of these. This reduces the number of cross-terms and higher-order terms that must be estimated using local coefficients in the global model. The increase in global model accuracy is offset by the increased communication cost associated with transmitting the local models of the terms in addition to those of the features. It must be emphasized that the accuracy of the method is not directly dependent on the number of partitions but rather on the number of cross-terms.

To summarize the wavelet-based CDM-MR method presented here

1. Calculate the wavelet basis coefficients for the features or terms in each partition,
2. Use thresholding to select a subset of the largest absolute value coefficients for each feature, creating a minimum square error model of the local features,
3. Transmit the local coefficient models to a central site,
4. Generate estimates of wavelet coefficients for cross-terms and/or higher order terms directly from the local model coefficients, producing a global wavelet model,
5. Calculate regression coefficients directly from the global wavelet coefficient model.

4.3. Application and Benchmarking of CDM Regression

In this section we apply the CDM-MR technique to two benchmark data sets that have been analyzed by others using standard MR techniques. The purpose is to demonstrate the specifics of the CDM-MR technique on tractable data sets and to compare the parametric regression models obtained with published results for centralized data techniques.

We begin with data first proposed in [29] and currently included in the NIST Statistical Reference Datasets [31]. This data set consists of values of Total employment, GNP implicit price deflater, GNP, Unemployment, Armed Forces manpower level, Non-institutionalized population $\geq$ 14 yrs., and Year, for the years 1947 through 1962. The first of these seven variables (Employment) is taken as the dependent variable and the data set is used to fit a MR model with the six independent variables and an intercept. We deal with the intercept

by adding a dummy feature to the problem. This feature has value 1 for all samples and the associated regression coefficient in the wavelet basis will correspond to the intercept value in the original basis.

The wavelet-packet decomposition for the "Year" independent variable is shown in Figure 4. The top ($j = 4$) row is simply the sample data itself that is assumed to represent the coefficients s_0^4 through s_{15}^4 of the scaling functions. Since we are using Haar wavelets ($H = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}), G = (\frac{1}{\sqrt{2}}, \frac{-1}{\sqrt{2}})$) the first entry in the $j = 3$ scale space is

$$s_0^3 = \frac{s_0^4}{\sqrt{2}} + \frac{s_1^4}{\sqrt{2}} = \frac{1947}{\sqrt{2}} + \frac{1948}{\sqrt{2}} = \frac{3895}{\sqrt{2}}$$

Likewise, the first entry for the $j = 3$ wavelet space is

$$d_0^3 = \frac{s_0^4}{\sqrt{2}} - \frac{s_1^4}{\sqrt{2}} = \frac{1947}{\sqrt{2}} - \frac{1948}{\sqrt{2}} = \frac{-1}{\sqrt{2}}$$

The second entry in each of the $j = 3$ spaces is

$$s_1^3 = \frac{s_2^4}{\sqrt{2}} + \frac{s_3^4}{\sqrt{2}} \text{ and } d_1^3 = \frac{s_2^4}{\sqrt{2}} - \frac{s_3^4}{\sqrt{2}}$$

with the other entries determined likewise. This process is repeated $log_2 16 = 4$ times, each time starting with the scale and wavelet spaces is the previous step to produce the complete wavelet-packet transform as shown in Figure 4. The complete set of $j = 0$ coefficients for the dependent variable, the six independent variables, and the dummy feature variable are shown in Table 2.

The results obtained by performing a regression on the complete set of wavelet coefficients is presented in Table 3. These results are, as expected, equivalent to the Certified Regression Statistics provided by NIST. Since all of the wavelet coefficients are used in the regression a centralized regression is being performed. Statistics for MR models created using 12, 8 and 5 wavelet coefficients per feature are shown in Tables 4, 5 and 6, respectively. These results show the model parameters and statistics change as the number of wavelet coefficients retained per feature approaches the minimum number required to produce an $\mathbf{X}$ matrix of proper rank. It should be noted that the effect of eliminating a small number of coefficients for this example is pronounced due to the small size of the data set and the relatively high collinearity among the independent variables in the Longley data set. However, these models represent the minimum square error models for the number of wavelet coefficients per feature retained. The implications of this may be seen by comparing the CDM-MR model generated from the eight largest wavelet coefficients for each feature (Table 5) and a centralized model generated from some combination of eight examples from the original data set. Tables 7 and 8 present the statistics for models generated with data samples for years 1947, 1949, 1953, 1954, 1955, 1956, 1960 and 1962, and for years 1947, 1948, 1949, 1950, 1959, 1960, 1961 and 1962. The statistics of these three models show some variation but the estimated regression coefficients for each model fall within the 95% confidence interval for the other two models. Of particular interest is the inter-model variation in the size of the 95% confidence intervals. The second centralized regression model has a much tighter confidence interval that the first centralized model and the CDM-MR model has a tighter confidence interval that either of the centralized data models. This result demonstrates that selecting some subset of the sample data to

centralize and then use to generate a regression model may not generate the MSE model for that amount of information transfer. The result also supports the MSE model result for the wavelet basis.

The second benchmark data set we employ is the Boston Housing data set created by Harrison and Rubinfeld [18]. This data set consists of 506 samples with 13 independent variables, 12 of which are real-valued, and one real-valued dependent variable. Since the regression model reported in [18] has an intercept term we add one dummy variable with a value of 1.0 for each sample to this data set. Descriptions of the variables are provided in Table 9. The regression model fit to the data is

$$\begin{aligned} log(MV) = \quad & a_1 + a_2 RM^2 + a_3 AGE + a_4 log(DIS) + a_5 log(RAD) \\ + \; & a_6 TAX + a_7 PTRATIO + a_8 (B - 0.63)^2 + a_9 log(LSTAT) \\ + \; & a_{10} CRIM + a_{11} ZN + a_{12} INDUS + a_{13} CHAS + a_{14} NOX^2 \end{aligned}$$

The Longley data provided a good example of how overall model statistics compare between CDM-MR models and models created using centralized data. However, the data set is to small to examine how the model changes as the number of wavelet coefficients per feature retained to build the model is reduced. Using the Boston Housing data set we compare the model produced with centralized data to models produced using 50%, 30% and 10% of the coefficients per feature. Table 10 presents the results of the comparison. The centralized data results are taken from [18] except for the R^2 value which was calculated using from the given coefficients and the data set. Overall, the CDM-MR model coefficients remain consistent with the centralized model coefficients even at the 10% retained wavelet coefficients per feature level.

These data sets were selected to facilitate demonstration of the CDM-MR technique. They are not typical of the data sets that CDM is intended for in that they are small enough to be stored at a single location and are complete in the sense that all features are under the direct control of the investigator. The data sets CDM targets are very large with different features sets residing at different locations and possibly under the control of different entities. Inevitably, interesting data sets that fit this description are proprietary. In order to evaluated the performance of CDM-MR for larger data sets we turn in the following section to the use of synthetic data sets.

4.4. CDM Regression Method Performance Trends

In this section we use several large (up to 100 MB) synthetic data sets to provide a characterization of CDM-MR method performance trends in terms of 1) appropriate selection of wavelet functions, 2) the number of cross-terms and higher-order terms relative to the number of features, 3) the sample size for a given problem. In addition we demonstrate scalability.

The basic metric we use to measure this performance is the residual value o_i defined by

$$y_i - \hat{y}_i = o_i$$

where y_i is the actual function or dependent feature value for example i and $\hat{y}_i$ is the estimate of that value generated by the regression model. This is similar to the residual value calculated in classical parametric regression but has the important difference that the examples used are from out-of-sample data not the examples used to build the regression

$j=4$	1947	1948	1949	1950	1951	1952	1953	1954	1955	1956	1957	1958	1959	1960	1961	1962
$j=3$	$3895/\sqrt{2}$	$3899/\sqrt{2}$	$3903/\sqrt{2}$	$3907/\sqrt{2}$	$3911/\sqrt{2}$	$3915/\sqrt{2}$	$3919/\sqrt{2}$	$3923/\sqrt{2}$	$-1/\sqrt{2}$	$-1/\sqrt{2}$	$-1/\sqrt{2}$	$-1/\sqrt{2}$	$-1/\sqrt{2}$	$-1/\sqrt{2}$	$-1/\sqrt{2}$	$-1/\sqrt{2}$
$j=2$	3897	3905	3913	3921	-2	-2	-2	-2	-1	-1	-1	-1	0	0	0	0
$j=1$	$7802/\sqrt{2}$	$7834/\sqrt{2}$	$-8/\sqrt{2}$	$-8/\sqrt{2}$	$-4/\sqrt{2}$	$-4/\sqrt{2}$	0	0	$-2/\sqrt{2}$	$-2/\sqrt{2}$	0	0	0	0	0	0
$j=0$	7818	-16	-8	0	-4	0	0	0	-2	0	0	0	0	0	0	0

FIG. 4. Wavelet-packet decomposition of Year variable from Longley data.

TABLE 2

Wavelet coefficients (j = 0) for Longley data.

Filter	Y	X1	X2	X3	X4	X5	X6	X7
HHHH	261268.0	4.0	406.725	1550793.75	12773.25	10426.75	469696.0	7818.0
GHHH	-12061.5	0.0	-35.475	-330374.25	-2176.75	-475.75	-22896.0	-16.0
HGHH	-5587.5	0.0	-19.625	-180627.75	-244.75	-1600.75	-11966.5	-8.0
GGHH	-816.0	0.0	-2.425	-3423.75	1362.25	-2038.25	2108.5	0.0
HHGH	-1068.0	0.0	-7.525	-64972.25	-1682.25	-71.75	-6475.5	-4.0
GHGH	166.5	0.0	2.075	14241.75	-202.25	-140.25	1260.5	0.0
HGGH	708.5	0.0	-2.275	5170.25	-439.25	196.75	1108.0	0.0
GGGH	280.0	0.0	1.525	-3377.75	-28.25	-204.75	-355.0	0.0
HHHG	-829.0	4.0	-4.575	-36884.75	-540.25	-15.75	-2895.5	-2.0
GHHG	326.5	0.0	-0.275	3168.25	-49.25	-83.25	645.5	0.0
HGHG	-169.5	0.0	-1.725	-700.25	-110.25	241.75	380.0	0.0
GGHG	-1143.0	0.0	-0.225	-17265.25	1061.75	-42.75	-310.0	0.0
HHGG	-1153.0	0.0	-1.625	-5499.75	621.25	-50.25	112.0	0.0
GHGG	438.5	0.0	-0.925	-3944.75	166.25	-211.75	-81.0	0.0
HGGG	-485.5	0.0	-0.975	-3763.25	142.25	149.25	33.5	0.0
GGGG	1417.0	0.0	-0.675	14615.75	-1229.75	280.75	67.5	0.0

model. The average residual value over the test data set

$$\bar{o} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)$$

will remain close to zero unless the CDM-MR method introduces a bias or offset into the model. The standard deviation of the residual values

$$\sigma_0 = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(o_i - \bar{o}_i)}$$

provides a notion of the distribution of the residuals about the mean. As we are interested only in relative performance of the models the standard deviation of residual values for a set of models that are to be compared are normalized such that the maximum standard deviation is 1.

Wavelets techniques form a large and expanding body of knowledge [19]. There are many families of wavelet functions. Any set of orthogonal wavelet functions may potentially be used in wavelet-based CDM. An important consideration in selecting a specific wavelet function for a problem is how sparse the feature representation becomes in terms of the wavelet coefficients. As the representation becomes sparser fewer wavelet coefficients are needed for a local model of a given accuracy. For this paper Haar wavelets were selected because their relatively simple mathematical form provides a necessary clarity in the development of the distributed MR technique. The Haar wavelets are the least-smooth members of the wavelet families they belong to.

TABLE 3

CDM Regression model statistics for Longley data with all coefficients retained.

R^2	0.999987				
$\hat{\sigma}$	304.8541				
	Coefficient	$\hat{\sigma}_i$	$t_{0.25}$	P	95% CI
$\hat{b}_1$	-3482258.7	890420.4	-3.9108	0.0036	-5496531.0, -1467986.3
$\hat{b}_2$	15.0619	84.9149	0.1774	0.8631	-177.0292, 207.1529
$\hat{b}_3$	-0.0358	0.0335	-1.0695	0.3127	-0.1116, 0.0399
$\hat{b}_4$	-2.0202	0.4884	-4.1364	0.0025	-3.1251, -0.9154
$\hat{b}_5$	-1.0332	0.2143	-4.8220	0.0009	-1.5179, -0.5485
$\hat{b}_6$	-0.0511	0.2261	-0.2261	0.8262	-0.5625, 0.4603
$\hat{b}_7$	1829.1515	455.4785	4.0159	0.0030	798.7867, 2859.5162

TABLE 4

CDM Regression model statistics for Longley data with 12 coefficients retained.

R^2	0.999985				
$\hat{\sigma}$	332.3977				
	Coefficient	$\hat{\sigma}_i$	$t_{0.25}$	P	95% CI
$\hat{b}_1$	-3459035.5	896750.4	-3.8573	0.0039	-5487627.4, -1430443.5
$\hat{b}_2$	56.1487	89.4239	0.6279	0.5457	-146.1424, 258.4398
$\hat{b}_3$	-0.0449	0.0316	-1.4186	0.1897	-0.1164, 0.0267
$\hat{b}_4$	-2.1178	0.4699	-4.5073	0.0015	-3.1808, -1.0549
$\hat{b}_5$	-1.0881	0.2429	-4.4794	0.0015	-1.6376, -0.5386
$\hat{b}_6$	0.0373	0.2254	0.1653	0.8723	-0.4726, 0.5472
$\hat{b}_7$	1811.8521	460.4716	3.9348	0.0034	770.1922, 2853.5119

TABLE 5

CDM Regression model statistics for Longley data with 8 coefficients retained.

R^2	0.999968				
$\hat{\sigma}$	477.9735				
	Coefficient	$\hat{\sigma}_i$	$t_{0.25}$	P	95% CI
$\hat{b}_1$	-2806889.6	1081972.7	-2.5942	0.0290	-5254483.9, -359295.4
$\hat{b}_2$	-24.3543	213.7620	-0.1139	0.9118	-507.9179, 459.2093
$\hat{b}_3$	-0.0216	0.0482	-0.4472	0.6653	-0.1306, 0.0875
$\hat{b}_4$	-1.7310	0.7048	-2.4561	0.0364	-3.3253, -0.1367
$\hat{b}_5$	-0.8998	0.4533	-1.9849	0.0784	-1.9253, 0.1257
$\hat{b}_6$	0.0081	0.3599	0.0225	0.9826	-0.8060, 0.8222
$\hat{b}_7$	1478.6216	556.3516	2.6577	0.0261	220.0666, 2737.1767

TABLE 6

CDM Regression model statistics for Longley data with 5 coefficients retained.

R^2	0.999941
$\hat{\sigma}$	650.4082

	Coefficient	$\hat{\sigma}_i$	$t_{0.25}$	P	95% CI
$\hat{b}_1$	-3958376.7	3597783.7	-1.1002	0.2998	12097135.1, 4180381.6
$\hat{b}_2$	-1975.3736	1285.5059	-1.5367	0.1588	-4883.3922, 932.6450
$\hat{b}_3$	0.1719	0.1223	1.4055	0.1934	-0.1048, 0.4487
$\hat{b}_4$	-1.2519	0.6300	-1.9871	0.0782	-2.6771, 0.1733
$\hat{b}_5$	-0.4363	0.5620	-0.7765	0.4574	-1.7076, 0.8349
$\hat{b}_6$	-0.2650	1.3296	-0.1993	0.8464	-3.2729, 2.7428
$\hat{b}_7$	2145.8926	1896.6631	1.1314	0.2871	-2144.6608, 6436.4460

TABLE 7

Standard regression model statistics for Longley data with 1947, 1948, 1949, 1950, 1959, 1960, 1961 and 1962 retained.

R^2	0.999665
$\hat{\sigma}$	191.1414

	Coefficient	$\hat{\sigma}_i$	$t_{0.25}$	P	95% CI
$\hat{b}_1$	-3401168.5	667110.9	-5.0984	0.1233	-11877579.8, 5075242.8
$\hat{b}_2$	-304.4729	211.3409	-1.4407	0.3863	-989.8016, 2380.8557
$\hat{b}_3$	0.0473	0.0455	1.0410	0.4872	-0.5305, 0.6252
$\hat{b}_4$	-0.9499	0.7564	-1.2558	0.4281	-10.5609, 8.6612
$\hat{b}_5$	-0.7709	0.4938	-1.5611	0.3627	-7.0456, 5.5038
$\hat{b}_6$	-0.8696	0.3721	-2.3370	0.2574	-5.5974, 3.8582
$\hat{b}_7$	1834.9335	342.7404	5.3537	0.1176	-2519.9775, 6189.8445

TABLE 8

Standard regression model statistics for Longley data with 1947, 1949, 1953, 1954, 1955, 1956, 1960 and 1962 retained.

R^2	0.998917
$\hat{\sigma}$	414.2000

	Coefficient	$\hat{\sigma}_i$	$t_{0.25}$	P	95% CI
$\hat{b}_1$	-4951145.8	5312775.9	-0.9319	0.5224	-72456074.4, 62553782.9
$\hat{b}_2$	276.7046	266.4325	1.0386	0.4880	-3108.6267, 3662.0359
$\hat{b}_3$	-0.1459	0.1829	-0.7978	0.5713	-2.4704, 2.1785
$\hat{b}_4$	-3.1734	2.6451	-1.1997	0.4424	-36.7822, 30.4355
$\hat{b}_5$	0.7333	2.5085	0.2923	0.8190	-31.1405, 32.6070
$\hat{b}_6$	0.6293	0.8046	0.7822	0.5774	-9.5938, 10.8525
$\hat{b}_7$	2548.0656	2719.4141	0.9401	0.5196	31890.8638, 36986.9950

TABLE 9

Variable definitions for Boston housing data set.

Variable	Description
MV	Median value of owner occupied homes.
RM	Average number of rooms per dwelling.
AGE	Proportion of owner-occupied units built prior to 1940.
B	Proportion of population that is black.
LSTAT	% of lower status of the population.
CRIM	Per capita crime rate.
ZN	Proportion of residential lad zoned for lots over 25,000 sq.ft.
INDUS	Proportion of non-retail business acres per town.
TAX	Full-value property-tax rate per $10,000.
PTRATIO	Pupil-teacher ratio by town.
DIS	Weighted distance to five Boston employment centers.
RAD	Index of accessibility to radial highways.
NOX	Nitric oxides concentration.

TABLE 10

Regression model results for Boston housing data set.

Parameter	Centralized Model	CDM Model - 50%	CDM Model - 30%	CDM Model - 10%
R^2	0.80569	0.804251	0.792491	0.775627
$\hat{a}_1$	9.76	9.73	9.72	9.67
$\hat{a}_2$	0.0063	0.0074	0.0090	0.0112
$\hat{a}_3$	0.0000898	0.0004	-0.00079	-0.00067
$\hat{a}_4$	-0.19	-0.20	-0.23	-0.19
$\hat{a}_5$	0.096	0.097	0.082	0.075
$\hat{a}_6$	-0.00042	-0.00042	-0.00033	-0.00039
$\hat{a}_7$	-0.031	-0.030	-0.027	-0.025
$\hat{a}_8$	0.36	0.37	0.36	0.36
$\hat{a}_9$	-0.37	-0.36	-0.34	-0.28
$\hat{a}_{10}$	-0.012	-0.013	-0.013	-0.014
$\hat{a}_{11}$	0.0000803	0.00032	0.00064	0.0016
$\hat{a}_{12}$	0.000241	0.00031	0.00102	0.0016
$\hat{a}_{13}$	0.088	0.084	0.119	0.118
$\hat{a}_{14}$	-0.0064	-0.0066	-0.0057	-0.0052

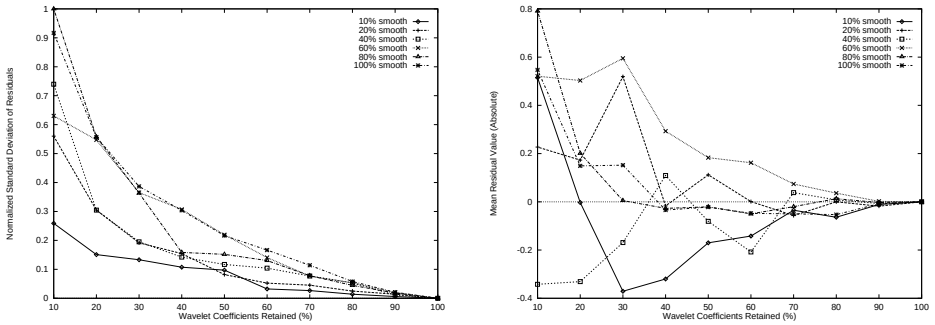

FIG. 5. Model accuracy loss due to retention of fewer wavelet coefficients is reduced as the feature sample characteristics become more consistent with the smoothness of the wavelet function.

To demonstrate the importance of matching the wavelet function to the data characteristics we turn the problem on its head and apply the Haar based distributed MR to a series of data sets which provide varying degrees of suitability for use with the Haar wavelets. A function with 15 linear terms and four cross-terms (19 total terms) based of a feature set of size 15 was used for this purpose. Feature samples were generated randomly with uniform probability on the interval $[-10.0, 10.0]$. The "smoothness" of the data was varied by introducing a probability that the sample value of each feature would change form one sample to the next. The smoothest data was generated with a 100% probability that feature sample values would change from one sample to the next (barring the chance that the random number generator returned the same value). The least smooth data had only a 10% probability of individual sample values changing from one sample to the next. Test cases were performed with smoothness between 10% and 100% to evaluate the impact on global model accuracy as a function of percentage of wavelet coefficients retained in the local models. Figure 5 shows that for a given percentage of retained wavelet coefficients the less-smooth data sets tend to be more accurate since they are more compatible with the wavelet functions. The plots also show that global model accuracy, as measured by normalized residual value standard deviation and mean residual value increases as the percentage of wavelet coefficients retained increases.

In Figure 6 the effect of varying numbers of cross-terms relative to a fixed number of features is shown. The base function used in this evaluation was the same one used to evaluate the effect of data smoothness with additional cross-terms added as required by the case. The feature sample data was 10% smooth and the local models used 10% wavelet coefficient retention. Figure 6 shows a linear relation between the number of cross-terms and normalized residual standard deviation. The mean residual value remains close to zero.

The importance of sample size in CDM was described in Section 3 of this paper. The effect of sample size was evaluated using a 36-term quadratic in 15 features [25]. In the function each feature is represented in a linear and a quadratic term and there are six additional cross-terms,

$$\begin{aligned} f(\mathbf{x}) \;=\; & 20x_0 + 5x_0^2 + 18x_1 - 8x_1^2 + 16x_2 + 13x_2^2 + \\ & 14x_3 + 11x_3^2 + 12x_4 - 14x_4^2 + 10x_5 - 8x_5^2 + \end{aligned}$$

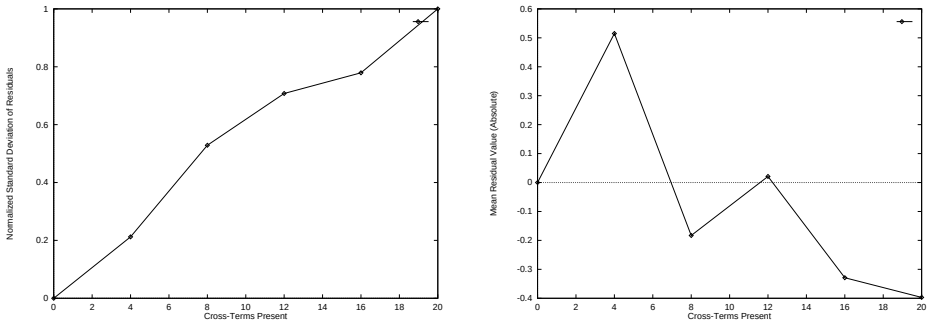

FIG. 6. Residual value standard deviation increases linearly while the mean residual value remains near zero as the number of cross-terms increase

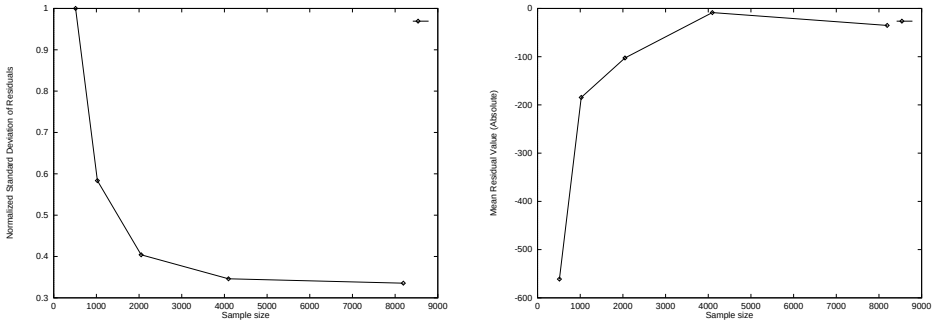

FIG. 7. Global model accuracy increases with sample size

$$8x_6 + 11x_6^2 + 6x_7 + 13x_7^2 - 7x_8 - 12x_8^2 -$$
$$9x_9 + 15x_9^2 - 11x_{10} + 9x_{10}^2 - 13x_{11} - 10x_{11}^2 -$$
$$15x_{12} - 16x_{12}^2 - 17x_{13} - 10x_{13}^2 - 19x_{14} + 7x_{14}^2 +$$
$$5x_0x_1 - 3x_6x_{10} + 12x_4x_{11} - 8x_{12}x_{14} - 2x_{13}x_2^2 + 4x_4^2x_8$$

Data sets were created with 10% smoothness and between 1K and 8K training examples and an equivalent number of test examples. Local models retained 10% of the wavelet coefficients. Figure 7 shows that global model accuracy does increase as the sample size increases.

To evaluate scalability of the CDM-MR algorithm a data set consisting of 193 features (1 dependent and 192 independent) and 64K samples was constructed. The dependent feature was a function of 128 total terms, 64 linear terms and 64 cross-terms. The data set represents time-series data for three types of process. One third of the features are similar to the " less smooth" data used in the other synthetic data sets. For this portion of the data set the features take on values in $[0, 1]$ and change from sample to sample with probability 0.1%. The second third of the features evolve according to the process

$$dS = (m - \frac{1}{2}\sigma^2)Sdt + \sigma Sdw$$

where dw is a Wiener process [20]. The sample data values are generated according to

$$S_{t+1} = S_t exp^{((m-\frac{1}{2}\sigma^2)\Delta t + \sigma\sqrt{\Delta t}Z)}$$

TABLE 11
Large data set calculations with various portions of coefficients retained.

% Coefficients Retained	Mean residual	Normalized standard deviation
100%	0.006	0.252
50%	-0.015	0.281
30 %	-0.007	0.401
10%	0.023	1.000

where Z is a random draw from a standard normal distribution and the specific values used are $m = 0.05$, $\sigma = 0.20$ and $S_0 = 1.0$. The final third of the features evolve according to the process

$$dr = a(b - r)dt + \sigma dw$$

where again dw is a Wiener process [20]. The sample data values are generated according to

$$r_{t+1} = r_t + a(b - r_t)\Delta t + \sigma\sqrt{\Delta t}Z$$

where Z is a random draw from a standard normal distribution, r_0 is set randomly on $[0.06, 0.1]$, $b = r_0$ and $a = 0.1$. For the second and third process $\Delta t = \frac{1}{2048}$.

The 64 linear terms in the regression function are based on the first third of the features and the 64 cross-terms are based on the products of the second and final third of the features. The function value itself is the sum of the 128 terms plus a normally distributed random error. Calculations were performed with 100% , 50% , 30% and 10% of the wavelet coefficients for each feature retained. The results of the calculations are presented in Table 11. The results show that for very large data sets the accuracy of the global model, as measured by the mean residual value and normalized standard deviation of residuals does not degrade significantly as the proportion of wavelet coefficients retained per feature is reduced to 10%.

It should be noted that it is currently rare to use parametric regression directly for a function with this many terms. Principal components and/or factor analysis are typically used in this situation.

4.5. CDM Regression Algorithm Performance

Given N samples of each feature the wavelet-packet decomposition algorithm performs N calculations for each of $log_2 N$ decomposition levels resulting in a time complexity of $O(NlogN)$. If M features reside in a partition then it will take $O(MNlogN)$ time to calculated the coefficients for all the features in that partition.

Once all retained wavelet coefficients have been centralized it may be necessary to calculate cross-term coefficients. Assuming that there are N samples, that r percent of the wavelet coefficients for each feature retained, and that the cross-terms depends on features in at most p partitions, then the worst-case performance will be $O((rN)^p)$.

The regression algorithm as implemented for this work is dominated by the time needed to set up the simultaneous equations. Given L independent regressors including the dummy variable for the intercept term if needed the setup requires $O(NL^2)$ time.

5. LINEAR DISCRIMINANT ANALYSIS

Linear discriminant analysis (LDA) [13] is another form of supervised learning that is related to MR. Given two populations for which the same features are measured samples from each population with known membership are used to construct a decision rule. Observations with unknown population membership may be correctly classified with high probability using the decision rule.

5.1. Linear Discriminant Analysis in CDM

An equivalence between MR and LDA was pointed out by Fisher [12]. Within the regression model pseudo-variables representing the population classes are employed as the dependent variables. For a two-class problem:

$$f(\bar{x}_i) = \begin{cases} c_1 & \text{if observation } i \text{ comes from class 1} \\ c_2 & \text{if observation } i \text{ comes from class 2} \end{cases}$$

Fisher proposed that the values of the pseudo-variables be

$$c_1 = \frac{n_2}{n_1 + n_2}; c_2 = \frac{-n_1}{n_1 + n_2} \tag{9}$$

where n_1 and n_2 are the number of training examples from class 1 and 2, respectively.

From a theoretical standpoint the difference between MR and Fisher's LDA is that in the case of MR the independent variables are assumed to be known exactly with any variability embodied in the dependent variable while in LDA the dependent variable (class) is known exactly and any variability is embodied in the independent variables. From an implementation standpoint the important difference between MR and LDA is how the model is used, not the basic CDM or regression techniques used to create the model. In the case of LDA the result of applying the regression model to a set of features in an observation of unknown class is compared to a decision boundary value that is determined as part of the learning or training phase. By using the pseudo-variable values proposed by Fisher for class values and with the assumption that the feature covariances are not significantly different between populations the decision boundary value becomes 0.0. What constitutes a significant difference in feature covariance between populations within the CDM framework and how this may be evaluated for any given populations is left for future resolution.

The following section describes an application of LDA to the Iris data set [12] a widely used benchmark data set for statistics and machine learning.

5.2. CDM Linear Discriminant Analysis Example: Iris Data

In this section distributed LDA applied to the Iris data set [12] that consists of measurements of four features of three varieties of Iris flower. The data set contains 150 examples, 50 for each variety or class. Two samples were randomly eliminated from each class and the remaining 144 were divided into three groups of 48 samples each, 16 from each class, in order to facilitate a 3-fold cross validation of the model. For the purposes of this demonstration each feature is assumed to reside in a separate partition and the class label column vector is only needed to generate the global models so it is not transmitted to each site. Since the Iris data represents three classes, not two, an additional step is required in the modeling process. First regression models that discriminate between each pair of classes are generated then those models were used in committee form to select the proper classification for observations of unknown class. In the case of a tie (each model selects a

different class) the estimated class value closest to the assigned class value for any model is used to select the classification.

The results of the validation test cases are shown in Table 12. On average the models created using wavelet-based CDM correctly classified the out-of-sample test cases 90.3% of the time. Examples of reported accuracies for centralized methods are presented in Table 13.

TABLE 12

Results of 3-fold cross-validation on Iris data.

Case	Correct Classification	Incorrect Classification	Accuracy
1	48	0	100.0%
2	41	7	84.0%
3	41	7	84.0%
Combined	130	14	90.3%

It should be noted that because there are no cross-terms or higher-order terms only four wavelet coefficients were required from each partition in order to generate the global model. Thus the communication overhead was sightly over 4% of that required to centralize all 96 sample values in any one training set at one site. Further, for this problem the communication cost of four wavelet coefficients is independent of the sample size. This is a result of the high level of compatibility between the discrete representation of the class variable and the Haar wavelet functions.

6. CONCLUSIONS AND FUTURE WORK

This paper presents a method for performing distributed multivariate regression using wavelet-based Collective Data Mining. The distributed multivariate parametric regression technique presented here learns local information in terms of the coefficients of an orthogonal basis function representation, transmits a small (relative to the sample size) number of significant coefficients to a central site and then generates a global model directly from that small set of significant coefficients. The method seamlessly blends machine learning and the theory of communication with the statistical methods employed in multivariate parametric regression to provide an effective data mining technique for use in a distributed data and computation environment.

In application to distributed multivariate parametric regression wavelet techniques were shown to produce an orthogonal basis that provides a sparse, distributed, representation of a function as basis function coefficients. Using these coefficients to communicate local

TABLE 13

Reported classification method accuracy for Iris data.

Method	Accuracy	Source
Clustering (ISODATA)	88.0%	(Freemen, 1970) [14]
Reduced-NN	93.0 - 96.7%	(Gates, 1972) [15]
1-NN	95 - 97%	(Duda & Hart, 1973) [11]
Partitional Clustering	89.337%	(Duda & Hart, 1973) [11]
Hierarchical Clustering	90.0%	(Duda & Hart, 1973) [11]
Tree	97.33%	(Duda & Hart, 1973) [11]

model information to a central site required as little as 10% of the communication cost required to assemble a centralized data set. The importance of selecting wavelet functions that are compatible with data characteristics, the reduction in model accuracy as the relative number of non-linear cross terms increases, and the increase in model accuracy with sample size were demonstrated.

Application of wavelet-based CDM methodology to linear discriminant analysis, a technique related to multivariate regression, was also presented. An application to the Iris data set with the assumption that each feature resides in a different data base showed classification accuracy similar to centralized techniques. Linear discriminant problems such as Iris are particularly well suited for treatment with the Haar wavelets used in this work due to the discrete nature of the class feature. Communication costs for this problem were shown to be directly proportional to the number of independent features in the discriminant function and independent of the sample size.

Future work will follow two distinct paths, further exploration of the use of wavelet techniques in this context and extension of these CDM techniques to other real-domain learning problems.

The work presented in this paper is based on Haar wavelets and the Haar-Walsh wavelet packet basis. Higher order (smoother) orthogonal wavelets and other wavelet bases, such as multi-resolution or Paley order, may provide improved performance in some cases. The ability to pre-characterize data sets in terms of appropriate wavelet function basis is currently being investigated.

Work on extending the real-domain wavelet-based CDM techniques to learning neural network models is ongoing.

ACKNOWLEDGMENT

The authors would like to acknowledge that this work is partially supported by a grant from the American Cancer Society. The Boston housing data set used in Section 4 was obtained from the StatLib library maintained by Carnegie Mellon University. The Iris data set used in the linear discriminant analysis presented in Section 5 was obtained from the UCI repository of machine learning databases [2].

REFERENCES

1. J. M. Aronis, V. Kolluri, F. J. Provost, and B. G. Buchanan. The world: Knowledge discovery from multiple distributed data bases. Technical Report ISL-96-6, Intelligent Systems Laboratory, Department of Computer Science, University of Pittsburgh, Pittsburgh, PA, 1996.
2. E. Blake, E. Keogh, and C.J. Merz. U.C.I. Repository of machine learning databases, 1998.
3. R. Carmona, Wen-Liang Hwang, and Bruno Torresani. *Practical Time-Frequency Analysis*, volume 9. Academic Press, San Diego, 1998.
4. B. Carnahan, H. A. Luther, and J. O. Wilkes. *Applied numerical methods*. John Wiley and Sons, Inc., New York, 1969.
5. P. Chan and S. Stolfo. Experiments on multistrategy learning by meta-learning. In *Proceeding of the Second International Conference on Information Knowledge Management*, pages 314–323, 1993.
6. P. Chan and S. Stolfo. Toward parallel and distributed learning by meta-learning. In *Working Notes AAAI Work. Knowledge Discovery in Databases*, pages 227–240. AAAI, 1993.
7. P. Chan and S. Stolfo. Toward scalable learning with non-uniform class and cost distribution: A case study in credit card fraud detection. In *Proceeding of the Fourth International Conference on Knowledge Discovery and Data Mining*, page Not available. AAAI Press, September 1998.
8. J. Chattratichat, J. Darlington, Y. Guo, S. Hedvall, M. Kohler, A. Saleem, J Sutiwaraphun, and D. Yang. Toward scalable learning with non-uniform class and cost distribution: A case study in credit card fraud detection. In *Proceeding of the Fourth International Conference on Knowledge Discovery and Data Mining*, page Not available. AAAI Press, September 1998.

9. D. Cheung, V. Ng, A. Fu, and Y. Fu. Efficient mining of association rules in distributed databases. *IEEE Transaction on Knowledge and Data Engineering*, 8(6):911–922, 1996.
10. V. Cho and B Wüthrich. Toward real time discovery from distributed information sources. In Xingdong Wu, Ramamohanarao Kotagiri, and Kevin B. Korb, editors, *Research and Development in Knowledge Discovery and Data Mining*, number 1394 in Lecture Notes in Computer Science : Lecture Notes in Artificial Intelligence, pages 376–377, New York, 1998. Springer-Verlag. Second Pacific-Asia Conference,PAKKD-98, Melbourne, Australia , April 1998.
11. R. O. Duda and D. E. Hart. *Pattern classification and scene analysis*. John Wiley and Sons, New York, 1973.
12. R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188, 1936.
13. B. Flury and H. Riedwyl. *Multivariate Statistics A Practical Approach*. Chapman and Hall, New York, 1988.
14. J. J. Freemen. Experiments in discrimination and classification. *Patterns Recognition*, 1:207–218, 1970.
15. G. W. Gates. The reduced nearest neighbor rule. *IEEE Transactions on Information Theory*, 18(3):431–433, 1972.
16. R. Grossman, S. Bailey, S. Kasif, D. Mon, A. Ramu, and B. Malhi. The preliminary design of papyrus: A system for high performance, distributed data mining over clusters, meta-clusters and super-clusters. Fourth International Conference of Knowledge Discovery and Data Mining, New York, New York, Pages 37–43, 1998.
17. H.F. Harmuth. *Transmission of Information by Orthogonal Functions*. Springer-Verlag, New York, 1972.
18. D. Harrison and D. L. Rubinfeld. Hedonic prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5:81–102, 1978.
19. Barbara Burke Hubbard. *The World According to Wavelets*. A. K. Peters, Ltd., Wellesley, MA, 1998.
20. John C. Hull. *Option, Futures, and Other Derivatives*. Prentice Hall, Upper Saddle River, NJ, 1997.
21. H. Kargupta. Distributed knowledge discovery: A brief overview. In *The Proceedings of the Spring 1999 Symposium of the Institute for operations Research and the Management Sciences (INFORMS)*, May 1999.
22. H. Kargupta, I. Hamzaoglu, and B. Stafford. Scalable, distributed data mining using an agent based architecture. In David Heckerman, Heikki Mannila, Daryl Pregibon, and Ramasamy Uthurusamy, editors, *Proceedings of Knowledge Discovery And Data Mining*, pages 211–214, Menlo Park, CA, 1997. AAAI Press.
23. H. Kargupta, I. Hamzaoglu, B. Stafford, V. Hanagandi, and K. Buescher. PADMA: Parallel data mining agent for scalable text classification. In *Proceedings Conference on High Performance Computing '97*, pages 290–295. The Society for Computer Simulation International, 1996.
24. H. Kargupta, E. Johnson, E. Riva Sanseverino, H. Park, L. D. Silvestre, and D. Hershberger. Scalable data mining from distributed, heterogeneous data, using collective learning and gene expression based genetic algorithms. Technical Report EECS-98-001, School of Electrical Engineering and Computer Science, Washington State University, 1998.
25. H. Kargupta, B. Park, D.E. Hershberger, and E. Johnson. Collective data mining: a new perspective toward distributed data mining. Technical Report EECS-99-001, Washington State University, Department of Electrical Engineering and Computer Science, 1999. To be published in the book "Advances in Distibuted and Parallel Knowledge Discovery." Eds: Hillol Kargupta and Philip Chan.
26. S. Kushilevitz and Y. Mansour. Learning decision rees using fourier spectrum. In *Proc. 23rd Annual ACM Symp. on Theory of Computing*, pages 455–464, 1991.
27. W. Lam and A. M. Segre. Distributed data mining of probabilistic knowledge. In *Proceedings of the 17th International Conference on Distributed Computing Systems*, pages 178–185, Washington, 1997. IEEE Computer Society Press.
28. W. Lee, S. Stolfo, and K. Mok. A data mining framework for adaptive intrusion detection. To appear in the Proceedings of the 1999 IEEE Symposium on Security and Privacy, IEEE Computer Society Press, 1999.
29. J. W. Longley. An appraisal of least squares programs for the electronic computer from the viewpoint of the user. *Journal of the American Statistical Association*, 62:819–841, 1967.
30. Frederick Mosteller and John W. Tukey. *Data Analysis and Regression*. Addison-Wesley, Menlo Park, CA, 1977.
31. N.I.S.T. Statistical reference datasets. `http:\\www.nist.gov\itl\div898\strd\`. Dataset Archives – Linear Regression – Longley.
32. F.J. Provost and K. Venkateswarlu. A survey of methods for scaling up inductive learning algorithms. *Data Mining and Knowledge Discovery*, 3(2):131–169, June 1999.

33. S. Stolfo et al. Jam: Java agents for meta-learning over distributed databases. In David Heckerman, Heikki Mannila, Daryl Pregibon, and Ramasamy Uthurusamy, editors, *Proceedings Third International Conference on Knowledge Discovery and Data Mining*, pages 74–81, Menlo Park, CA, 1997. AAAI Press.

34. Eric Stollnitz, Tony D. DeRose, and David H. Salesin. Wavelets for computer graphics: A primer, part 1. *IEEE Computer Graphics and Applications*, 5(3):76–84, May 1995.

35. Eric Stollnitz, Tony D. DeRose, and David H. Salesin. Wavelets for computer graphics: A primer, part 2. *IEEE Computer Graphics and Applications*, 5(4):75–85, June 1995.

36. R. Subramonian and S. Parthasarathy. An architecture for distributed data mining. Fourth International Conference of Knowledge Discovery and Data Mining, New York, New York, Pages 44–59, 1998.

37. M. V. Wickerhauser. *Adapted Wavelet Analysis from Theory to Software*. A. K. Peters Ltd., 1994.

38. D. Wolpert. Stacked generalization. *Neural Networks*, 5:241–259, 1992.

39. K. Yamanishi. Distributed cooperative bayesian learning strategies. In *Proceedings of COLT 97*, pages 250–262, New York, 1997. ACM.