A Survey on Wavelet Applications in Data Mining

Tao Li
Department of Computer Science
Univ. of Rochester
Rochester, NY 14627
taoli@cs.rochester.edu

Qi Li
Dept. of Computer & Information Sciences
Univ. of Delaware
Newark, DE 19716
qili@cis.udel.edu

Shenghuo Zhu
Department of Computer Science
Univ. of Rochester
Rochester, NY 14627
zsh@cs.rochester.edu

Mitsunori Ogihara
Department of Computer Science
Univ. of Rochester
Rochester, NY 14627
ogihara@cs.rochester.edu

ABSTRACT

Recently there has been significant development in the use of wavelet methods in various data mining processes. However, there has been written no comprehensive survey available on the topic. The goal of this is paper to fill the void. First, the paper presents a high-level data-mining framework that reduces the overall process into smaller components. Then applications of wavelets for each component are reviewd. The paper concludes by discussing the impact of wavelets on data mining research and outlining potential future research directions and applications.

1. INTRODUCTION

The wavelet transform is a synthesis of ideas that emerged over many years from different fields, such as mathematics and signal processing. Generally speaking, the wavelet transform is a tool that divides up data, functions, or operators into different frequency components and then studies each component with a resolution matched to its scale [52]. Therefore, the wavelet transform is anticipated to provide economical and informative mathematical representation of many objects of interest [1]. Nowadays many computer software packages contain fast and efficient algorithms to perform wavelet transforms. Due to such easy accessibility wavelets have quickly gained popularity among scientists and engineers, both in theoretical research and in applications. Above all, wavelets have been widely applied in such computer science research areas as image processing, computer vision, network management, and data mining.

Over the past decade data mining, or knowledge discovery in databases (KDD), has become a significant area both in academia and in industry. Data mining is a process of automatic extraction of novel, useful and understandable patterns from a large collection of data. Wavelet theory could naturally play an important role in data mining since it is well founded and of very practical use. Wavelets have many favorable properties, such as vanishing moments, hierarchical and multiresolution decomposition structure, linear time and space complexity of the transformations, decorrelated coefficients, and a wide variety of basis functions. These properties could provide considerably more efficient and effective solutions to many data mining problems. First, wavelets could provide presentations of data that make the mining process more efficient and accurate. Second, wavelets could be incorporated into the kernel of many data mining algorithms. Although standard wavelet applications are mainly on data which have temporal/spatial localities (e.g. time series, stream data, and image data) wavelets have also been successfully applied to diverse domains in data mining. In practice, a wide variety of wavelet-related methods have been applied to a wide range of data mining problems.

Although wavelets have attracted much attention in the data mining community, there has been no comprehensive review of wavelet applications in data mining. In this paper we attempt to fill the void by presenting the necessary mathematical foundations for understanding and using wavelets as well as a summary of research in wavelets applications. To appeal to a broader audience in the data mining community, this paper also providea brief overview of the practical research areas in data mining where wavelet could be used. The reader should be cautioned, however, that the wavelet is so a large research area that truly comprehensive surverys are almost impossible, and thus, that our overview may be a little eclectic. An interested reader is encouraged to consult with other papers for further reading, in particular, surveys of wavelet applicaations in statistics [1; 10; 12; 121; 127; 163], time series analysis [124; 44; 129; 121; 122], biological data [9], signal processing [110; 158], image processing [133; 115; 85] and others [117; 174]. Also, [93] provides a good overview on wavelet applications in database projects. The reader should be cautioned also that in our presentation mathematical descriptions are modified so that they adapt to data mining problems. A reader wishing to learn more mathematical details of wavelets is referred to [150; 52; 46; 116; 169; 165; 151].

This paper is organized as follows: To discuss a wide spectrum of wavelet applications in data mining in a systematic manner it seems crucial that data mining processes are divided into smaller components. Section 2 presents a high-level data mining framework, which reduces data mining process into four components. Section 3 introduces some necessary mathematical background related to wavelets. Then wavelet applications in each of the four components will be reviewed in Sections 4, 5, and 6. Section 7 discusses some other wavelet applications which are related to data mining. Finally, Section 8 discusses future research directions.

2. A FRAMEWORK FOR DATA MINING PROCESS

In this section, we give a high-level framework for data mining process and try to divide the data mining process into components. The purpose of the framework is to make our following reviews on wavelet applications in a more systematic way and hence it is colored to suit our discussion. More detailed treatment of the data mining process could be found in [79; 77].

Data mining or knowledge discovery is the nontrivial extraction of implicit, previously unknown, and potentially useful information from large collection of data. It can be viewed as a multidisciplinary activity because it exploits several research disciplines of artificial intelligence such as machine learning, pattern recognition, expert systems, knowledge acquisition, as well as mathematical disciplines such as statistics, information theory and uncertain inference. In our understanding, knowledge discovery refers to the overall process of extracting high-level knowledge from low-level data in the context of large databases. In the proposed framework, we view that knowledge discovery process usually consists of an iterative sequence of the following steps: **data management**, **data preprocessing**, **data mining tasks algorithms** and **post-processing**. These four steps are the four components of our framework.

First, **data management** concerns the specific mechanism and structures for how the data are accessed, stored and managed. The data management is greatly related to the implementation of data mining systems. Though many research papers do not elaborate explicit data management, it should be note that data management can be extremely important in practical implementations.

Next, **data preprocessing** is an important step to ensure the data quality and to improve the efficiency and ease of the mining process. Real-world data tend to be incomplete, noisy, inconsistent, high dimensional and multi-sensory etc. and hence are not directly suitable for mining. Data preprocessing usually includes data cleaning to remove noisy data and outliers, data integration to integrate data from multiple information sources, data reduction to reduce the dimensionality and complexity of the data and data transformation to convert the data into suitable forms for mining etc.

Third, we refer **data mining tasks and algorithms** as an essential step of knowledge discovery where various algorithms are applied to perform the data mining tasks. There are many different data mining tasks such as visualization, classification, clustering, regression and content retrieval etc. Various algorithms have been used to carry out these tasks and many algorithms such as Neural Network and Principal Component Analysis could be applied to several different kinds of tasks.

Finally, we need **post-processing** [28] stage to refine and evaluate the knowledge derived from our mining procedure. For example, one may need to simplify the extracted knowledge. Also, we may want to evaluate the extracted knowledge, visualize it, or merely document it for the end user. We may interpret the knowledge and incorporate it into an existing system, and check for potential conflicts with previously induced knowledge.

The four-component framework above provides us with a simple systematic language for understanding the steps that make up the data mining process. Since **post-processing** mainly concerns the non-technical work such as documentation and evaluation, we then focus our attentions on the first three components and will review wavelet applications in these components.

It should be pointed out that categorizing a specific wavelet technique/paper into a component of the framework is not strict or unique. Many techniques could be categorized as performing on different components. In this survey, we try to discuss the wavelet techniques with respect to the most relevant component based on our knowledge. When there is an overlap, i.e., a wavelet technique might be related to different components, we usually briefly examine the relationships and differences.

3. WAVELET BACKGROUND

In this section, we will present the basic foundations that are necessary to understand and use wavelets. A wavelet can own many attractable properties, including the essential properties such as compact support, vanishing moments and dilating relation and other preferred properties such as smoothness and being a generator of an orthonormal basis of function spaces $L^2(R^n)$ etc. Briefly speaking, compact support guarantees the localization of wavelets (In other words, processing a region of data with wavelets does not affect the the data out of this region); vanishing moment guarantees wavelet processing can distinguish the essential information from non-essential information; and dilating relation leads fast wavelet algorithms. It is the requirements of localization, hierarchical representation and manipulation, feature selection, and efficiency in many tasks in data mining that make wavelets be a very powerful tool. The other properties such as smoothness and generators of orthonormal basis are preferred rather than essential. For example, Haar wavelet is the simplest wavelet which is discontinuous, while all other Daubechies wavelets are continuous. Furthermore all Daubechies wavelets are generators of orthogonal basis for $L^2(R^n)$, while spline wavelets generate unconditional basis rather than orthonormal basis [47], and some wavelets could only generate redundant frames rather than a basis [138; 53]. The question that in what kinds of applications we should use orthonormal basis, or other (say unconditional basis, or frame) is yet to be solved. In this section, to give readers a relatively comprehensive view of wavelets, we will use Daubechies wavelets as our concrete examples. That is, in this survey, a wavelet we use is always assumed to be a generator of orthogonal basis.

In signal processing fields, people usually thought wavelets to be convolution filters which has some specially properties such as quadrature mirror filters (QMF) and high pass etc. We agree that it is convenient to apply wavelets to practical applications if we thought wavelets to be convolution filters. However, according to our experience, thinking of wavelets as functions which own some special properties such as compact support, vanishing moments and multiscaling etc., and making use of some simple concepts of function spaces $L^2(R^n)$ (such as orthonormal basis, subspace and inner product etc.) may bring readers a clear understanding why these basic properties of wavelets can be successfully applied in data mining and how these properties of wavelets may be applied to other problems in data mining. Thus in most uses of this survey, we treat wavelets as functions. In real algorithm designs and implementations, usually a function is straightforwardly discretized and treated as a vector. The interested readers could refer to [109] for more details on treating wavelets as filters.

The rest of the section is organized to help readers answer the fundamental questions about wavelets such as: what is a wavelet, why we need wavelets, how to find wavelets, how to compute wavelet transforms and what are the properties of wavelets etc. We hope readers could get a basic understanding about wavelet after reading this section.

3.1 Basics of Wavelet in $L^2(R)$

So, first, **what is a wavelet?** Simply speaking, a mother wavelet is a function $\psi(x)$ such that $\{\psi(2^j x - k), i, k \in Z\}$ is an orthonormal basis of $L^2(R)$. The basis functions are usually referred as wavelets[1]. The term wavelet means a small wave. The smallness refers to the condition that we desire that the function is of finite length or compactly supported. The wave refers to the condition that the function is oscillatory. The term mother implies that the functions with different regions of support that are used in the transformation process are derived by dilation and translation of the mother wavelet.

[1] A more formal definition of wavelet can be found in Appendix A. Note that this orthogonality is not an essential property of wavelets. We include it in the definition because we discuss wavelet in the context of Daubechies wavelet and orthogonality is a good property in many applications.

At first glance, wavelet transforms are pretty much the same as Fourier transforms except they have different bases. So **why bother to have wavelets? What are the real differences between them**? The simple answer is that wavelet transform is capable of providing time and frequency localizations simultaneously while Fourier transforms could only provide frequency representations. Fourier transforms are designed for stationary signals because they are expanded as sine and cosine waves which extend in time forever, if the representation has a certain frequency content at one time, it will have the same content for all time. Hence Fourier transform is not suitable for non-stationary signal where the signal has time varying frequency [130]. Since FT doesn't work for non-stationary signal, researchers have developed a revised version of Fourier transform, The Short Time Fourier Transform(STFT). In STFT, the signal is divided into small segments where the signal on each of these segments could be assumed as stationary. Although STFT could provide a time-frequency representation of the signal, Heisenberg's Uncertainty Principle makes the choice of the segment length a big problem for STFT. The principle states that one cannot know the exact time-frequency representation of a signal and one can only know the time intervals in which certain bands of frequencies exist. So for STFT, longer length of the segments gives better frequency resolution and poorer time resolution while shorter segments lead to better time resolution but poorer frequency resolution. Another serious problem with STFT is that there is no inverse, i.e., the original signal can not be reconstructed from the time-frequency map or the spectrogram.

Wavelet is designed to give good time resolution and poor frequency resolution at high frequencies and good frequency resolution and poor time resolution at low frequencies [130]. This is useful for many practical signals since they usually have high frequency components for a short durations (bursts) and low frequency components for long durations (trends). The time-frequency cell structures for STFT and WT are shown in Figure 1 and Figure 2 respectively.

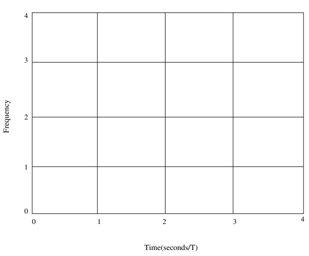

Figure 1: Time-Frequency structure of STFT. The graph shows that time and frequency localizations are independent. The cells are always square.

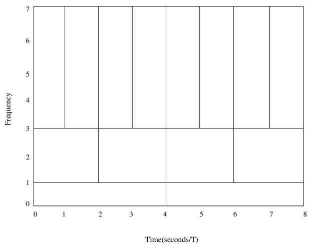

Figure 2: Time Frequency structure of WT. The graph shows that frequency resolution is good for low frequency and time resolution is good at high frequencies.

In data mining practice, the key concept in use of wavelets is the discrete wavelet transform(DWT). So our following discussion on wavelet is focused on discrete wavelet transform.

3.2 Dilation Equation

How to find the wavelets? The key idea is self-similarity. Start with a function $\phi(x)$ that is made up of smaller version of itself. This is the refinement (or 2-scale,dilation) equation

$$\phi(x) = \sum_{k=-\infty}^{\infty} a_k \phi(2x-k) \tag{3.1}$$

$a_k' s$ are called filter coefficients or masks. The function $\phi(x)$ is called the scaling function (or father wavelet). Under certain conditions,

$$\psi(x) = \sum_{k=-\infty}^{\infty} (-1)^k b_k \phi(2x-k) = \sum_{k=-\infty}^{\infty} (-1)^k \bar{a}_{1-k} \phi(2x-k) \tag{3.2}$$

gives a wavelet[2].

What are the conditions? First, the scaling function is chosen to preserve its area under each iteration, so $\int_{-\infty}^{\infty} \phi(x)dx = 1$. Integrating the refinement equation then

$$\int_{-\infty}^{\infty} \phi(x)dx = \sum a_k \int_{-\infty}^{-\infty} \phi(2x-k)dx = \frac{1}{2}\sum a_k \int_{-\infty}^{\infty} \phi(u)du$$

Hence $\sum a_k = 2$. So the stability of the iteration forces a condition on the coefficient a_k. Second, the convergence of wavelet expansion [3] requires the condition $\sum_{k=0}^{N-1} (-1)^k k^m a_k = 0$ where $m = 0, 1, 2, \ldots, \frac{N}{2} - 1$ (if a finite sum of wavelets is to represent the signal as accurately as possible). Third, requiring the orthogonality of wavelets forces the condition $\sum_{k=0}^{N-1} a_k a_{k+2m} = 0$ where $m = 0, 1, 2, \ldots, \frac{N}{2} - 1$. Finally if the scaling function is required to be orthogonal $\sum_{k=0}^{N-1} a_k^2 = 2$. To summarize, the conditions are

$$\begin{cases} \sum_{k=0}^{N-1} a_k = 2 & \text{stability} \\ \sum_{k=0}^{N-1} (-1)^k k^m a_k = 0 & \text{convergence} \\ \sum_{k=0}^{N-1} a_k a_{k+2m} = 0 & \text{orthogonality of wavelets} \\ \sum_{k=0}^{N-1} a_k^2 = 2 & \text{orthogonality of scaling functions} \end{cases}$$

This class of wavelet function is constrained, by definition, to be zero outside of a small interval. This makes the property of compact support. Most wavelet functions, when plotted, appear to be extremely irregular. This is due to the fact that the refinement equation assures that a wavelet $\psi(x)$ function is non-differentiable everywhere. The functions which are normally used for performing transforms consist of a few sets of well-chosen coefficients resulting in a function which has a discernible shape.

Let's now illustrate how to generate Haar[4] and Daubechies wavelets. They are named for pioneers in wavelet theory [75; 51]. First, consider the above constraints on the a_k for $N = 2$. The stability condition enforces $a_0 + a_1 = 2$, the accuracy condition implies $a_0 - a_1 = 0$ and the orthogonality gives $a_0^2 + a_1^2 = 2$. The unique solution is $a_0 = a_1 = 1$. if $a_0 = a_1 = 1$, then $\phi(x) = \phi(2x) + \phi(2x-1)$. The refinement function is satisfied by a box function

$$B(x) = \begin{cases} 1 & 0 \le x < 1 \\ 0 & \text{otherwise} \end{cases}$$

Once the box function is chosen as the scaling function, we then get the simplest wavelet: Haar wavelet, as shown in Figure 3.

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

[2] $\bar{a}$ means the conjugate of a. When a is a real number, $\bar{a} = a$.

[3] This is also known as the vanishing moments property.

[4] Haar wavelet represents the same wavelet as Daubechies wavelets with support at $[0, 1]$, called db_1.

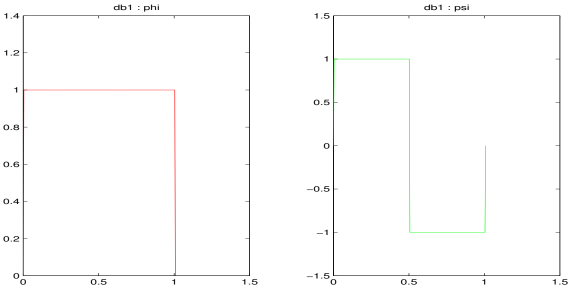

Figure 3: Haar Wavelet

Second, if $N = 4$, The equations for the masks are:

$$\begin{aligned} a_0 + a_1 + a_2 + a_3 &= 2 \\ a_0 - a_1 + a_2 - a_3 &= 0 \\ -a_1 + 2a_2 - 3a_3 &= 0 \\ a_0 a_2 + a_1 a_3 &= 0 \\ a_0^2 + a_1^2 + a_2^2 + a_3^2 &= 2 \end{aligned}$$

The solutions are $a_0 = \frac{1+\sqrt{3}}{4}, a_1 = \frac{3+\sqrt{3}}{4}, a_2 = \frac{3-\sqrt{3}}{4}, a_3 = \frac{1-\sqrt{3}}{4}$. The corresponding wavelet is Daubechies-2(db_2) wavelet that is supported on intervals $[0, 3]$, as shown in Figure 4. This construction is known as Daubechies wavelet construction [51]. In general, db_n represents the family of Daubechies Wavelets and n is the order. The family includes Haar wavelet since Haar wavelet represents the same wavelet as db_1. Generally it can be shown that

- The support for db_n is on the interval $[0, 2n - 1]$.
- The wavelet db_n has n vanishing moments[5].
- The regularity increases with the order. db_n has rn continuous derivatives (r is about 0.2).

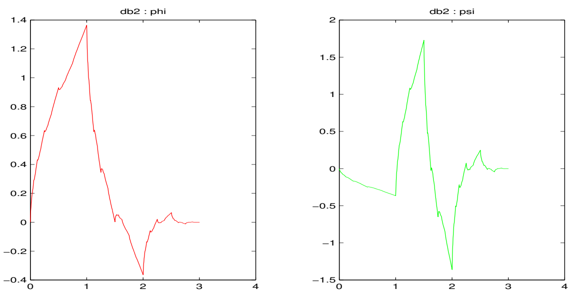

Figure 4: Daubechies-2(db_2) Wavelet

Finally let's look at some examples where the orthogonal property does not hold. If $a_{-1} = \frac{1}{2}, a_0 = 1, a_1 = \frac{1}{2}$, then

$$\phi(x) = \frac{1}{2}\phi(2x+1) + \phi(2x) + \frac{1}{2}\phi(2x-1).$$

The solution to this is the Hat function

$$\phi(x) = \begin{cases} x+1 & -1 \le x \le 0 \\ -(x-1) & 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

[5]We will discuss more about vanishing moments in Section 3.5.

So we would get $\psi(x) = -\frac{1}{2}\phi(2x+1) + \phi(2x) - \frac{1}{2}\phi(2x-1)$. Note that the wavelets generated by Hat function are not orthogonal. Similarly, if $a_{-2} = \frac{1}{8}, a_{-1} = \frac{1}{2}, a_0 = \frac{3}{4}, a_1 = \frac{1}{2}, a_2 = \frac{1}{8}$, we get cubic B-spline and the wavelets it generated are also not orthogonal.

3.3 Multiresolution Analysis(MRA) and fast DWT algorithm

How to compute wavelet transforms? To answer the question of efficiently computing wavelet transform, we need to touch on some material of MRA. Multiresolution analysis was first introduced in [102; 109] and there is a fast family of algorithms based on it [109]. The motivation of MRA is to use a sequence of embedded subspaces to approximate $L^2(R)$ so that people can choose a proper subspace for a specific application task to get a balance between accuracy and efficiency (Say, bigger subspaces can contribute better accuracy but waste computing resources). Mathematically, MRA studies the property of a sequence of closed subspaces $V_j, j \in Z$ which approximate $L^2(R)$ and satisfy

$$\cdots V_{-2} \subset V_{-1} \subset V_0 \subset V_1 \subset V_2 \subset \cdots,$$

$\overline{\bigcup_{j\in Z} V_j} = L^2(R)$ ($L^2(R)$ space is the closure of the union of all V_j) and $\bigcap_{j\in Z} V_j = \emptyset$ (the intersection of all V_j is empty). So **what does multiresolution mean?** The multiresolution is reflected by the additional requirement $f \in V_j \iff f(2x) \in V_{j+1}, j \in Z$ (This is equivalent to $f(x) \in V_0 \iff f(2^j x) \in V_j$),i.e., all the spaces are scaled versions of the central(reference) space V_0.

So **how does this related to wavelets**? Because the scaling function ϕ easily generates a sequence of subspaces which can provide a simple multiresolution analysis. First, the translations of $\phi(x)$, i.e., $\phi(x-k), k \in Z$, span a subspace, say V_0 (Actually, $\phi(x-k), k \in Z$ constitutes an orthonormal basis of the subspace V_0). Similarly $2^{-1/2}\phi(2x-k), k \in Z$ span another subspace, say V_1. The dilation equation 3.1 tells us that ϕ can be represented by a basis of V_1. It implies that ϕ falls into subspace V_1 and so the translations $\phi(x-k), k \in Z$ also fall into subspace V_1. Thus V_0 is embedded into V_1. With different dyadic, it is straightforward to obtain a sequence of embedded subspaces of $L^2(R)$ from only one function. It can be shown that the closure of the union of these subspaces is exactly $L^2(R)$ and their intersections are empty sets [52]. So here, we see that j controls the observation resolution while k controls the observation location.

Given two consecutive subspaces, say V_0 and V_1, it is natural for people to ask what information is contained in the complement space $V_1 \ominus V_0$, which is usually denoted as W_0. From equation 3.2, it is straightforward to see that ψ falls also into V_1 (and so its translations $\psi(x-k), k \in Z$). Notice that ψ is orthogonal to ϕ. It is easy to claim that an arbitrary translation of the father wavelet ϕ is orthogonal to an arbitrary translation of the mother wavelet ψ. Thus, the translations of the wavelet ψ span the complement subspace W_0. Similarly, for an arbitrary j, $\psi_{k,j}, k \in Z$, span an orthonormal basis of W_j which is the orthogonal complement space of V_j in V_{j+1}. Therefore, $L^2(R)$ space is decomposed into an infinite sequence of wavelet spaces, i.e., $L^2(R) = \bigoplus_{j\in Z} W_j$. More formal proof of wavelets' spanning complement spaces can be found in [52].

A direct application of multiresolution analysis is the fast discrete wavelet transform algorithm, called *pyramid* algorithm [109]. The core idea is to progressively smooth the data using an iterative procedure and keep the detail along the way, i.e., analyze projections of f to W_j. We use Haar wavelets to illustrate the idea through the following example. In Figure 5, the raw data is in resolution 3 (also called layer 3). After the first decomposition, the data are divided into two parts: one is of average information (projection in the scaling space V_2 and the other is of detail information (projection in the wavelet space W_2). We then repeat the similar decomposition on the data in V_2, and get the projection data in V_1 and W_1, etc. We also give a more formal treatment in Appendix B.

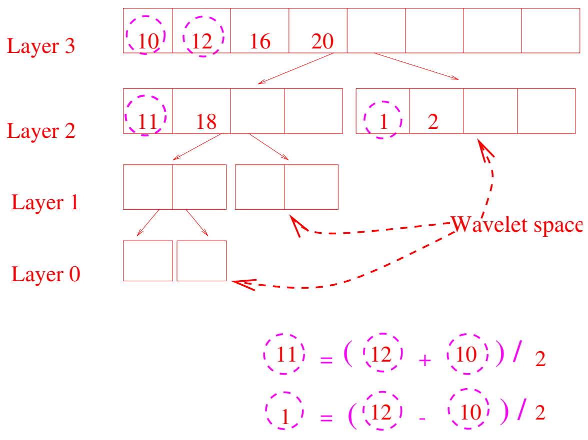

Figure 5: Fast Discrete Wavelet Transform

The fact that $L^2(R)$ is decomposed into an infinite wavelet subspace is equivalent to the statement that $\psi_{j,k}, j, k \in Z$ span an orthonormal basis of $L^2(R)$. An arbitrary function $f \in L^2(R)$ then can be expressed as follows:

$$f(x) = \sum_{j,k \in Z} d_{j,k}\psi_{j,k}(x), \tag{3.3}$$

where $d_{j,k} = \langle f, \psi_{j,k} \rangle$ is called *wavelet coefficients*. Note that j controls the observation resolution and k controls the observation location. If data in some location are relatively smooth (it can be represented by low-degree polynomials), then its corresponding wavelet coefficients will be fairly small by the vanishing moment property of wavelets.

3.4 Examples of Haar wavelet transform

In this section, we give two detailed examples of Haar wavelet transform.

3.4.1 One-dimensional transform

Haar transform can be viewed as a series of averaging and differencing operations on a discrete function. We compute the averages and differences between every two adjacent values of $f(x)$. The procedure to find the Haar transform of a discrete function $f(x)$ =[7 5 1 9] is shown in Table 1: Resolution 4 is the full resolution of the discrete function $f(x)$. In resolution 2, (6 5) are obtained by taking the average of (7 5) and (1 9) at resolution 4 respectively. (-1 4) are the differences of (7 5) and (1 9) divided by 2 respectively. This process is repeated until a resolution 1 is reached. The Haar transform $H(f(x))$ =(5.5 -0.5 -1 4) is obtained by combining the last average value 5 and the coefficients found on the right most column, -0.5, -1 and 4. In other words, the wavelet transform of original sequence is the single coefficient representing the overall average of the original average of the original numbers, followed by the detail coefficients in order of increasing resolutions. Different resolutions can be obtained by adding difference values back or subtracting differences from averages. For instance, (6 5)=(5.5+0.5,5.5−0.5) where 5.5 and −0.5 are the first and the second coefficient respectively. This process can be done recursively until the full resolution is reached. Note that no information has been gained or lost by this transform: the original sequence had 4 numbers and so does the transform.

Resolution	Approximations	Detail coefficients
4	7 5 1 9	
2	6 5	-1 4
1	5.5	-0.5

Table 1: An Example of One-dimensional Haar Wavelet Transform

Haar wavelets are the most commonly used wavelets in database/computer science literature because they are easy to comprehend and fast to compute. The *error tree* structure is often used by researchers in the field as a helpful tool for exploring and understanding the key properties of the Haar wavelets decomposition [113; 70]. Basically speaking, the *error tree* is a hierarchical structure built based on the wavelet decomposition process. The error tree of our example is shown in Figure 6. The leaves of the tree represents the original signal value and the internal nodes correspond to the wavelet coefficients. the wavelet coefficient associated with an internal node in the error tree contributes to the signal values at the leaves in its subtree. In particular, the root corresponds the overall average of the original data array. The depth of the tree represents the resolution level of the decomposition.

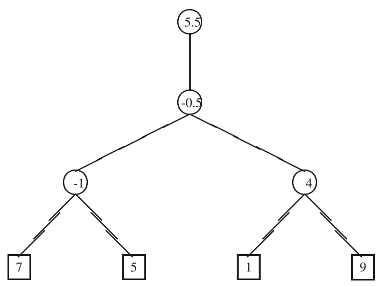

Figure 6: Error tree.

3.4.2 Multi-dimensional wavelet transform

Multi-dimensional wavelets are usually defined via the tensor product[6]. The two-dimensional wavelet basis consists of all possible tensor products of one-dimensional basis function[7]. In this section we will illustrate the two-dimensional Haar wavelet transform through the following example.

Let's compute the Haar wavelet transform of the following two-dimensional data

$$\begin{bmatrix} 3 & 5 & 6 & 7 \\ 9 & 8 & 7 & 4 \\ 6 & 5 & 7 & 9 \\ 4 & 6 & 3 & 8 \end{bmatrix}.$$

The computation is based on 2×2 matrices. Consider the upper left matrix

$$\begin{bmatrix} 3 & 5 \\ 9 & 8 \end{bmatrix}.$$

[6]For a given *component function* $f^1, \cdots f^d$, define $\prod_{j=1}^{d} f^j(x_1, \cdots, x_d) = \prod_{j=1}^{d} f^j(x_j)$ as the tensor product.

[7]There are also some non-standard constructions of high dimensional basis functions based on mutual transformations of the dimensions and interested readers may refer to [149] for more details.

We first compute the overall average: $(3+5+9+8)/4 = 6.25$, then the average of the difference of the summations of the rows: $1/2[(9+8)/2 - (3+5)/2] = 2.25$, followed by the average of the difference of the summations of the columns: $1/2[(5+8)/2 - (3+9)/2] = 0.25$ and finally the average of the difference of the summations of the diagonal: $1/2[(3+8)/2 - (9+5)/2] = -0.75$. So we get the following matrix

$$\begin{bmatrix} 6.25 & 2.25 \\ 0.25 & -0.75 \end{bmatrix}.$$

For bigger data matrices, we usually put the overall average element of all transformed 2×2 matrix into the first block, the average of the difference of the summations of the columns into the second block and so on. So the transformed matrix of the original data is

$$\begin{bmatrix} 6.25 & 6 & 2.25 & -0.50 \\ 5.25 & 6.75 & -0.25 & -1.25 \\ 0.25 & -0.50 & -0.75 & -1 \\ 0.25 & 1.75 & 0.75 & 0.75 \end{bmatrix}.$$

3.5 Properties of Wavelets

In this section, we summarize and highlight the properties of wavelets which make they are useful tools for data mining and many other applications. A wavelet transformation converts data from an original domain to a wavelet domain by expanding the raw data in an orthonormal basis generated by dilation and translation of a father and mother wavelet. For example, in image processing, the original domain is spatial domain, and the wavelet domain is frequency domain. An inverse wavelet transformation converts data back from the wavelet domain to the original domain. Without considering the truncation error of computers, the wavelet transformation and inverse wavelet transformation are lossless transformations. So the representations in the original domain and the wavelet domain are completely equivalent. In the other words, wavelet transformation preserves the structure of data. The properties of wavelets are described as follows:

1. Computation Complexity: First, the computation of wavelet transform can be very efficient. Discrete Fourier transform(DFT) requires $O(N^2)$ multiplications and fast Fourier transform also needs $O(N \log N)$ multiplications. However fast wavelet transform based on Mallat's pyramidal algorithm) only needs $O(N)$ multiplications. The space complexity is also linear.

2. Vanishing Moments: Another important property of wavelets is vanishing moments. A function $f(x)$ which is supported in bounded region ω is called to have n-vanishing moments if it satisfies the following equation:

$$\int_{\omega} f(x) x^j dx = 0, \quad j = 0, 1, \ldots, n. \qquad (3.4)$$

   That is, the integrals of the product of the function and low-degree polynomials are equal to zero. For example, Haar wavelet(or db_1) has 1-vanishing moment and db_2 has 2-vanishing moment. The intuition of vanishing moments of wavelets is the oscillatory nature which can thought to be the characterization of difference or details between a datum with the data in its neighborhood. Note that the filter [1, -1] corresponding to Haar wavelet is exactly a difference operator. With higher vanishing moments, if data can be represented by low-degree polynomials, their wavelet coefficients are equal to zero. So if data in some bounded region can be represented (approximated) by a low-degree polynomial, then its corresponding wavelet coefficient is (is close to) zero. Thus the vanishing moment property leads to many important wavelet techniques such as denoising and dimensionality reduction. The noisy data can usually be approximated by low-degree polynomial if the data are smooth in most of regions, therefore the corresponding wavelet coefficients are usually small which can be eliminated by setting a threshold.

3. Compact Support: Each wavelet basis function is supported on a finite interval. For example, the support of Haar function is [0,1]; the support of wavelet db_2 is [0, 3]. Compact support guarantees the localization of wavelets. In other words, processing a region of data with wavelet does not affect the the data out of this region.

4. Decorrelated Coefficients: Another important aspect of wavelets is their ability to reduce temporal correlation so that the correlation of wavelet coefficients are much smaller than the correlation of the corresponding temporal process [67; 91]. Hence, the wavelet transform could be able used to reduce the complex process in the time domain into a much simpler process in the wavelet domain.

5. Parseval's Theorem: Assume that $e \in L^2$ and ψ_i be the orthonormal basis of L^2. The Parseval's theorem states the following property of wavelet transform

$$\|e\|_2^2 = \sum_i |< e, \psi_i >|^2.$$

   In other words, the energy, which is defined to be the square of its L_2 norm, is preserved under the orthonormal wavelet transform. Hence the distances between any two objects are not changed by the transform.

In addition, the multiresolution property of scaling and wavelet functions, as we discussed in Section 3.3, leads to hierarchical representations and manipulations of the objects and has widespread applications. There are also some other favorable properties of wavelets such as the symmetry of scaling and wavelet functions, smoothness and the availability of many different wavelet basis functions etc. In summary, the large number of favorable wavelet properties make wavelets powerful tools for many practical problems.

4. DATA MANAGEMENT

One of the features that distinguish data mining from other types of data analytic tasks is the huge amount of data. So data management becomes very important for data mining. The purpose of data management is to find methods for storing data to facilitate fast and efficient access. Data management also plays an important role in the iterative and interactive nature of the overall data mining process. The wavelet transformation provides a natural hierarchy structure and multidimensional data representation and hence could be applied to data management.

Shahabi et al. [144; 143] introduced novel wavelet based tree structures: TSA-tree and 2D TSA-tree, to improve the efficiency of multilevel trends and surprise queries on time sequence data. Frequent queries on time series data are to identify rising and falling trends and abrupt changes at multiple level of abstractions. For example, we may be interested in the trends/surprises of the stock of Xerox Corporation within the last week, last month, last year or last decades. To support such multi-level queries, a large amount of raw data usually needs to be retrieved and processed. TSA (Trend and Surprise Abstraction) tree are designed to expedite the query

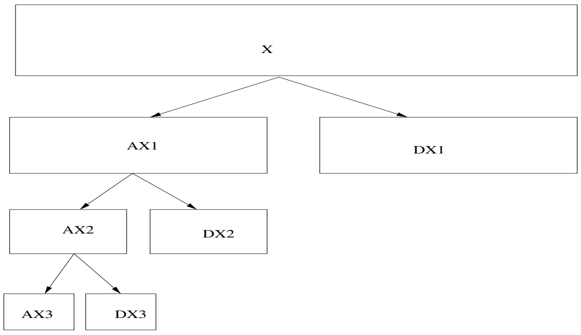

Figure 7: 1D TSA Tree Structure: X is the input sequence. AX_i and DX_i are the trend and surprise sequence at level i.

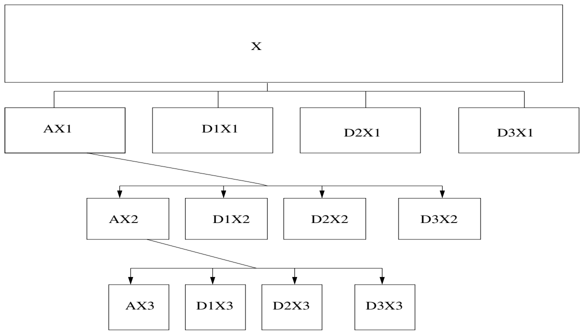

Figure 8: 2D TSA Tree Structure: X is the input sequence. $AX_i, D1X_i, D2X_i, D3X_i$ are the trend and horizontal, vertical and diagonal sequence at level i respectively.

process. TSA tree is constructed based on the procedure of discrete wavelet transform. The root is the original time series data. Each level of the tree corresponds to a step in wavelet decomposition. At the first decomposition level, the original data is decomposed into a low frequency part (trend) and a high frequency part (surprise). The left child of the root records the trend and the right child records the surprise. At the second decomposition level, the low frequency part obtained in the first level is further divided into a trend part and a surprise part. So the left child of the left child of the root records the new trend and the right child of the left child of the root records the new surprise. This process is repeated until the last level of the decomposition. The structure of the TSA tree is described in Figure 7. Hence as we traverse down the tree, we increase the level of abstraction on trends and surprises and the size of the node is decreased by a half. The nodes of the TSA tree thus record the trends and surprises at multiple abstraction levels. At first glance, TSA tree needs to store all the nodes. However, since TSA tree encodes the procedure of discrete wavelet transform and the transform is lossless, so we need only to store the all wavelet coefficients (i.e., all the leaf nodes). The internal nodes and the root can be easily obtained through the leaf nodes. So the space requirement is identical to the size of original data set. In [144], the authors also propose the techniques of dropping selective leaf nodes or coefficients with the heuristics of energy and precision to reduce the space requirement. 2D TSA tree is just the two dimensional extensions of the TSA tree using two dimensional discrete wavelet transform. In other words, the 1D wavelet transform is applied on the 2D data set in different dimensions/direction to obtain the trends and the surprises. The surprises at a given level correspond to three nodes which account for the changes in three different directions: horizontal, vertical and diagonal. The structure of a 2D TSA-tree is shown in Fig 8.

Venkatesan et al. [160] proposed a novel image indexing technique based on wavelets. With the popularization of digital images, managing image databases and indexing individual images become more and more difficult since extensive searching and image comparisons are expensive. The authors introduce an image hash function to manage the image database. First a wavelet decomposition of the image is computed and each subband is randomly tiled into small rectangles. Each rectangle's statistics (e.g., averages or variances) are calculated and quantized and then input into the decoding stage and a suitably chosen error-correcting code to generate the final hash value. Experiments have shown that the image hashing is robust against common image processing and malicious attacks.

Santini and Gupta [141] defined wavelet transforms as a data type for image databases and also presents an algebra to manipulate the wavelet data type. It also mentions that wavelets can be stored using a quadtree structure for every band and hence the operations can be implemented efficiently. Subramanya and Youssef [155] applied wavelets to index the Audio data. More wavelet applications for data management can be found in [140]. We will discuss more about image indexing and search in Section 6.5.

5. PREPROCESSING

Real world data sets are usually not directly suitable for performing data mining algorithms [134]. They contain noise, missing values and may be inconsistent. In addition, real world data sets tend to be too large, high-dimensional and so on. Therefore, we need data cleaning to remove noise, data reduction to reduce the dimensionality and complexity of the data and data transformation to convert the data into suitable form for mining etc. Wavelets provide a way to estimate the underlying function from the data. With the vanishing moment property of wavelets, we know that only some wavelet coefficients are significant in most cases. By retaining selective wavelet coefficients, wavelets transform could then be applied to denoising and dimensionality reduction. Moreover, since wavelet coefficients are generally decorrelated, we could transform the original data into wavelet domain and then carry out data mining tasks. There are also some other wavelet applications in data preprocessing. In this section, we will elaborate various applications of wavelets in data preprocessing.

5.1 Denoising

Noise is a random error or variance of a measured variable [78]. There are many possible reasons for noisy data, such as measurement/instrumental errors during the data acquisition, human and computer errors occurring at data entry, technology limitations and natural phenomena such as atmospheric disturbances, etc. Removing noise from data can be considered as a process of identifying outliers or constructing optimal estimates of unknown data from available noisy data. Various smoothing techniques, such as binning methods, clustering and outlier detection, have been used in data mining literature to remove noise. Binning methods smooth a sorted data value by consulting the values around it. Many data mining algorithms find outliers as a by-product of clustering algorithms [5; 72; 176] by defining outliers as points which do not lie in clusters. Some other techniques [87; 14; 135; 94; 25] directly find points which behave very differently from the normal ones. Aggarwal and Yu [6] presented new techniques for outlier detection by studying the behavior of projections from datasets. Data can also be smoothed by using regression methods to fit them with a function. In addition, the post-pruning techniques used in decision trees are able to avoid the overfitting problem caused by noisy data [119]. Most of these methods, however, are not specially designed to deal with noise and noise reduction and smoothing are only side-products of learning algorithms for other tasks. The information loss caused by these methods is also a problem.

Wavelet techniques provide an effective way to denoise and have been successfully applied in various areas especially in image research [39; 152; 63]. Formally, Suppose observation data $y = (y_1, \ldots, y_n)$ is a noisy realization of the signal $x = (x_1, \ldots, x_n)$:

$$y_i = x_i + \epsilon_i, \quad i = 1, \ldots, n, \tag{5.5}$$

where ϵ_i is noise. It is commonly assumed that ϵ_i are independent from the signal and are independent and identically distributed (*iid*) Gaussian random variables. A usual way to denoise is to find $\hat{x}$ such that it minimizes the mean square error (MSE),

$$MSE(\hat{x}) = \frac{1}{n}\sum_{i=1}^{n}(\hat{x}_i - x_i)^2. \tag{5.6}$$

The main idea of wavelet denoising is to transform the data into a different basis, the wavelet basis, where the *large* coefficients are mainly the useful information and the *smaller* ones represent noise. By suitably modifying the coefficients in the new basis, noise can be directly removed from the data.

Donoho and Johnstone [60] developed a methodology called *waveShrink* for estimating x. It has been widely applied in many applications and implemented in commercial software, e.g., wavelet toolbox of Matlab [69].

WaveShrink includes three steps:

1. Transform data y to the wavelet domain.
2. Shrink the empirical wavelet coefficients towards zero.
3. Transform the shrunk coefficients back to the data domain.

There are three commonly used shrinkage functions: the hard, soft and the non-negative garrote shrinkage functions:

$$\delta_\lambda^H(x) = \begin{cases} 0 & |x| \le \lambda \\ x & |x| > \lambda \end{cases}$$

$$\delta_\lambda^S(x) = \begin{cases} 0 & |x| \le \lambda \\ x - \lambda & x > \lambda \\ \lambda - x & x < -\lambda \end{cases}$$

$$\delta_\lambda^H(x) = \begin{cases} 0 & |x| \le \lambda \\ x - \lambda^2/x & |x| > \lambda \end{cases}$$

where $\lambda \in [0, \infty)$ is the threshold.

Wavelet denoising generally is different from traditional filtering approaches and it is nonlinear, due to a thresholding step. Determining threshold λ is the key issue in waveShrink denoising. Minimax threshold is one of commonly used thresholds. The *minimax*[8] *threshold* λ^* is defined as threshold λ which minimizes expression

$$\inf_\lambda \sup_\theta \left\{ \frac{R_\lambda(\theta)}{n^{-1} + \min(\theta^2, 1)} \right\}, \tag{5.7}$$

where $R_\lambda(\theta) = E(\delta_\lambda(x) - \theta)^2, x \sim N(\theta, 1)$. Interested readers can refer to [69] for other methods and we will also discuss more about the choice of threshold in Section 6.3. Li et al. [104] investigated the use of wavelet preprocessing to alleviate the effect of noisy data for biological data classification and showed that, if the localities of data the attributes are strong enough, wavelet denoising is able to improve the performance.

[8]Minimize Maximal Risk.

5.2 Data Transformation

A wide class of operations can be performed directly in the wavelet domain by operating on coefficients of the wavelet transforms of original data sets. Operating in the wavelet domain enables to perform these operations progressively in a coarse-to-fine fashion, to operate on different resolutions, manipulate features at different scales, and to localize the operation in both spatial and frequency domains. Performing such operations in the wavelet domain and then reconstructing the result is more efficient than performing the same operation in the standard direct fashion and reduces the memory footprint. In addition, wavelet transformations have the ability to reduce temporal correlation so that the correlation of wavelet coefficients are much smaller than the correlation of corresponding temporal process. Hence simple models which are insufficient in the original domain may be quite accurate in the wavelet domain. These motivates the wavelet applications for data transformation. In other words, instead of working on the original domain, we could working on the wavelet domain.

Feng et al. [65] proposed a new approach of applying Principal Component Analysis (PCA) on the wavelet subband. Wavelet transform is used to decompose an image into different frequency subbands and a mid-range frequency subband is used for PCA representation. The method reduces the computational load significantly while achieving good recognition accuracy. Buccigrossi and Simoncelli [29] developed a probability model for natural images, based on empirical observation of their statistics in the wavelet transform domain. They noted that pairs of wavelet coefficients, corresponding to basis functions at adjacent spatial locations, orientations, and scales, generally to be non-Gaussian in both their marginal and joint statistical properties and specifically, their marginals are heavy-tailed, and although they are typically decorrelated, their magnitudes are highly correlated. Hornby et al. [82] presented the analysis of potential field data in the wavelet domain. In fact, many other wavelet techniques that we will review for other components could also be regarded as data transformation.

5.3 Dimensionality Reduction

The goal of dimension reduction[9] is to express the original data set using some smaller set of data with or without a loss of information. Wavelet transformation represents the data as a sum of prototype functions and it has been shown that under certain conditions the transformation only related to selective coefficients. Hence similar to denoising, by retaining selective coefficients, wavelets can achieve dimensionality reduction. Dimensionality reduction can be thought as an extension of the data transformation presented in Section 5.2: while data transformation just transforms original data into wavelet domain without discarding any coefficients, dimensionality reduction only keeps a collection of selective wavelet coefficients.

More formally, the dimensionality reduction problem is to project the n-dimensional tuples that represent the data in a k-dimensional space so that $k << n$ and the distances are preserved as well as possible. Based on the different choices of wavelet coefficients, there are two different ways for dimensionality reduction using wavelet,

- Keep the largest k coefficients and approximate the rest with 0,
- Keep the first k coefficients and approximate the rest with 0.

[9]Some people also refer this as feature selection.

Keeping the largest k coefficients achieve more accurate representation while keeping the first k coefficients is useful for indexing [74]. Keeping the first k coefficients implicitly assumes a priori the significance of all wavelet coefficients in the first k coarsest levels and that all wavelet coefficients at a higher resolution levels are negligible. Such a strong prior assumption heavily depends on a suitable choice of k and essentially denies the possibility of local singularities in the underlying function [1].

It has been shown that [148; 149], if the basis is orthonormal, in terms of L_2 loss, maintaining the largest k wavelet coefficients provides the optimal k-term Haar approximation to the original signal. Suppose the original signal is given by $f(x) = \sum_{i=0}^{M-1} c_i \mu_i(x)$ where $\mu_i(x)$ is an orthonormal basis. In discrete form, the data can then be expressed by the coefficients $c_0, \cdots, c_{M-1}$. Let σ be a permutation of $0, \ldots, M-1$ and $f'(x)$ be a function that uses the first M' number of coefficients of permutation σ, i.e., $f'(x) = \sum_{i=0}^{M'-1} c_{\sigma(i)} \mu_{\sigma(i)}(x)$. It is then straightforward to show that the decreasing ordering of magnitude gives the best permutation as measured in L_2 norm. The square of L_2 error of the approximation is

$$
\begin{aligned}
&||f(x) - f'(x)||_2^2 \\
&= \left\langle f(x) - f'(x), f(x) - f'(x) \right\rangle \\
&= \left\langle \sum_{i=M'}^{M-1} c_{\sigma(i)} \mu_{\sigma(i)}, \sum_{j=M'}^{M-1} c_{\sigma(j)} \mu_{\sigma(j)} \right\rangle \\
&= \sum_{i=M'}^{M-1} \sum_{j=M'}^{M-1} c_{\sigma(i)} c_{\sigma(j)} \left\langle \mu_{\sigma(i)}, \mu_{\sigma(j)} \right\rangle = \sum_{i=M'}^{M-1} (c_{\sigma(i)})^2
\end{aligned}
$$

Hence to minimize the error for a given M', the best choice for σ is the permutation that sorts the coefficients in decreasing order of magnitude; i.e., $|c_{\sigma(0)}| \geq c_{\sigma(1)} \geq \cdots \geq c_{\sigma(M-1)}$.

Using the largest k wavelet coefficients, given a predefined precision ϵ, the general step for dimension reduction can be summarized in the following steps:

- Compute the wavelet coefficients of the original data set.
- Sort the coefficients in order of decreasing magnitude to produce the sequence $c_0, c_1, \ldots, c_{M-1}$.
- Starting with $M' = M$, find the best M' such that $\sum_{i=M'}^{M-1} ||c_i|| \leq \epsilon$.

$||c_i||$ is the norm of c_i. In general, the norm can be chosen as L_2 norm where $||c_i|| = (c_i)^2$ or L_1 norm where $||c_i|| = |c_i|$ or other norms. In practice, wavelets have been successfully applied in image compression [45; 37; 148] and it was suggested that L_1 norm is best suited for the task of image compression [55].

Chan and Fu [131] used the first k coefficients of Haar wavelet transform of the original time series for dimensionality reduction and they also show that no false dismissal (no qualified results will be rejected) for range query and nearest neighbor query by keeping the first few coefficients.

6. DATA MINING TASKS AND ALGORITHMS

Data mining tasks and algorithms refer to the essential procedure where intelligent methods are applied to extract useful information patterns. There are many data mining tasks such as clustering, classification, regression, content retrieval and visualization etc. Each task can be thought as a particular kind of problem to be solved by a data mining algorithm. Generally there are many different algorithms could serve the purpose of the same task. Meanwhile, some algorithms can be applied to different tasks. In this section, we review the wavelet applications in data mining tasks and algorithms. We basically organize the review according to different tasks. The tasks we discussed are clustering, classification, regression, distributed data mining, similarity search, query processing and visualization. Moreover, we also discuss the wavelet applications for two important algorithms: Neural Network and Principal/Independent Component Analysis since they could be applied to various mining tasks.

6.1 Clustering

The problem of clustering data arises in many disciplines and has a wide range of applications. Intuitively, the clustering problem can be described as follows: Let W be a set of n data points in a multi-dimensional space. Find a partition of W into classes such that the points within each class are *similar* to each other. The clustering problem has been studied extensively in machine learning [41; 66; 147; 177], databases [5; 72; 7; 73; 68], and statistics [22; 26] from various perspectives and with various approaches and focuses.

The multi-resolution property of wavelet transforms inspires the researchers to consider algorithms that could identify clusters at different scales. WaveCluster [145] is a multi-resolution clustering approach for very large spatial databases. Spatial data objects can be represented in an n-dimensional feature space and the numerical attributes of a spatial object can be represented by a feature vector where each element of the vector corresponds to one numerical attribute (feature). Partitioning the data space by a grid reduces the number of data objects while inducing only small errors. From a signal processing perspective, if the collection of objects in the feature space is viewed as an n-dimensional signal, the high frequency parts of the signal correspond to the regions of the feature space where there is a rapid change in the distribution of objects (i.e., the boundaries of clusters) and the low frequency parts of the n-dimensional signal which have high amplitude correspond to the areas of the feature space where the objects are concentrated (i.e., the clusters). Applying wavelet transform on a signal decomposes it into different frequency sub-bands. Hence to identify the clusters is then converted to find the connected components in the transformed feature space. Moreover, application of wavelet transformation to feature spaces provides multiresolution data representation and hence finding the connected components could be carried out at different resolution levels. In other words, the multi-resolution property of wavelet transforms enable the WaveCluster algorithm could effectively identify arbitrary shape clusters at different scales with different degrees of accuracy. Experiments have shown that WaveCluster outperforms Birch [176] and CLARANS [126] by a large margin and it is a stable and efficient clustering method.

6.2 Classification

Classification problems aim to identify the characteristics that indicate the group to which each instance belongs. Classification can be used both to understand the existing data and to predict how new instances will behave. Wavelets can be very useful for classification tasks. First, classification methods can be applied on the wavelet domain of the original data as discussed in Section 5.2 or selective dimensions of the wavelet domain as we will discussed in this section. Second, the multi-resolution property of wavelets can be incorporated into classification procedures to facilitate the process.

Castelli et al. [33; 34; 35] described a wavelet-based classification algorithm on large two-dimensional data sets typically large digital images. The image is viewed as a real-valued configuration on a rectangular subset of the integer lattice Z^2 and each point on the lattice (i.e. pixel) is associated with a vector denoting as pixel-values and a label denoting its class. The classification problem here consists of observing an image with known pixel-values but unknown labels and assigning a label to each point and it was motivated primarily by the need to classify quickly and efficiently large images in digital libraries. The typical approach [50] is the traditional pixel-by-pixel analysis which besides being fairly computationally expensive, also does not take into account the correlation between the labels of adjacent pixels. The wavelet-based classification method is based on the progressive classification [35] framework and the core idea is as follows: It uses generic (parametric or non-parametric) classifiers on a low-resolution representation of the data obtained using discrete wavelet transform. The wavelet transformation produce a multiresolution pyramid representation of the data. In this representation, at each level each coefficient corresponds to a $k \times k$ pixel block in the original image. At each step of the classification, the algorithm decides whether each coefficient corresponds to a homogeneous block of pixels and assigns the same class label to the whole block or to re-examine the data at a higher resolution level. And the same process is repeated iteratively. The wavelet-based classification method achieves a significant speedup over traditional pixel-wise classification methods. For images with pixel values that are highly correlated, the method will give more accurate results than the corresponding non-progressive classifier because DWT produces a weight average of the values for a $k \times k$ block and the algorithm tend to assume more uniformity in the image than may appear when we look at individual pixels. Castelli et al. [35] presented the experimental results illustrating the performance of the method on large satellite images and Castelli et al. [33] also presented theoretical analysis on the method.

Blume and Ballard [23] described a method for classifying image pixels based on learning vector quantization and localized Haar wavelet transform features. A Haar wavelet transform is utilized to generate a feature vector per image pixel and this provides information about the local brightness and color as well as about the texture of the surrounding area. Hand-labeled images are used to generated the a codebook using the optimal learning rate learning vector quantization algorithm. Experiments show that for small number of classes, the pixel classification is as high as 99%.

Scheunders et al. [142] elaborated texture analysis based on wavelet transformation. The multiresolution and orthogonal descriptions could play an important role in texture classification and image segmentation. Useful gray-level and color texture features can be extracted from the discrete wavelet transform and useful rotation-invariant features were found in continuous transforms. Sheikholeslami [146] presented a content-based retrieval approach that utilizes the texture features of geographical images. Various texture features are extracted using wavelet transforms. Using wavelet-based multi-resolution decomposition, two different sets of features are formulated for clustering. For each feature set, different distance measurement techniques are designed and experimented for clustering images in database. Experimental results demonstrate that the retrieval efficiency and effectiveness improve when the clustering approach is used. Mojsilovic et al. [120] also proposed a wavelet-based approach for classification of texture samples with small dimensions. The idea is first to decompose the given image with a filter bank derived from an orthonormal wavelet basis and to form an image approximation with nigher resolution. Texture energy measures calculated at each output of the filter bank as well as energies if synthesized images are used as texture features for a classification procedure based on modified statistical t-test.The new algorithm has advantages in classification of small and noisy samples and it represents a step toward structural analysis of weak textures. More usage on texture classification using wavelets can be found in [100; 40]. Tzanetakis et al. [157] used wavelet to extract a feature set for representing music surface and rhythm information to build automatic genre classification algorithms.

6.3 Regression

Regression uses existing values to forecast what other values will be and it is one of the fundamental tasks of data mining. Consider the standard univariate nonparametric regression setting: $y_i = g(t_i) + \epsilon_i, i = 1, \ldots, n$ where ϵ_i are independent $N(0, \sigma^2)$ random variables. The goal is to recover the underlying function g from the noisy data y_i, without assuming any particular parametric structure for g. The basic approach of using wavelets for nonparametric regression is to consider the unknown function g expanded as a generalized wavelet series and then to estimate the wavelet coefficients from the data. Hence the original nonparametric problem is thus transformed to a parametric one [1]. Note that the denoise problem we discussed in Section 5.1 can be regarded as a subtask of the regression problem since the estimation of the underlying function involves the noise removal from the observed data.

6.3.1 *Linear Regression*

For linear regression, we can express

$$g(t) = c_0\phi(t) + \sum_{j=0}^{\infty} \sum_{k=0}^{2^j-1} w_{jk}\psi_{jk}(t),$$

where $c_0 =< g, \phi >, w_{jk} =< g, \psi_{jk} >$. If we assume g belongs to a class of functions with certain regularity, then the corresponding norm of the sequence of w_{jk} is finite and w_{jk}'s decay to zero. So

$$g(t) = c_0\phi(t) + \sum_{j=0}^{M} \sum_{k=0}^{2^j-1} w_{jk}\psi_{jk}(t)$$

for some M and a corresponding truncated wavelet estimator is [1]

$$\hat{g}_M(t) = \hat{c}_0\phi(t) + \sum_{j=0}^{M} \sum_{k=0}^{2^j-1} \hat{w}_{jk}\psi_{jk}(t).$$

Thus the original nonparametric problem reduces to linear regression and the sample estimates of the coefficients are given by:

$$\hat{c}_0 = \frac{1}{n}\sum_{i=1}^{n} \phi(t_i)y_i, \hat{w}_{jk} = \frac{1}{n}\sum_{i=1}^{n} \psi_{jk}(t_i)y_i.$$

The performance of the truncated wavelet estimator clearly depends on an appropriate choice of M. Various methods such as Akaike's Information Criterion [8] and cross-validation can be used for choosing M. Antoniadis [11] suggested linear shrunk wavelet estimators where the $\hat{w_{jk}}$ are linearly shrunk by appropriately chosen level-dependent factors instead of truncation. We should point out that: the linear regression approach here is similar to the dimensionality reduction by keeping the first several wavelet coefficients discussed in section 5.3. There is an implicit strong assumption underlying the approach. That is, all wavelet coefficients in the first M coarsest levels are significant while all wavelet coefficients at a higher resolution levels are negligible. Such a strong assumption clearly would not hold for many functions. Donoho and Johnstone [60] showed that no linear estimator will be optimal in minimax sense for estimating inhomogeneous functions with local singularities. More discussion on linear regression can be found in [10].

6.3.2 Nonlinear Regression

Donoho et al. [58; 61; 60; 59] proposed a nonlinear wavelet estimator of g based on reconstruction from a more judicious selection of the empirical wavelet coefficients. The vanishing moments property of wavelets makes it reasonable to assume that essentially only a few 'large' $\hat{w}_{jk}$ contain information about the underlying function g, while 'small' $\hat{w}_{jk}$ can be attributed to noise. If we can decide which are the 'significant' large wavelet coefficients, then we can retain them and set all the others equal to zero, so obtaining an approximate wavelet representation of underlying function g.

The key concept here is thresholding. Thresholding allows the data itself to decide which wavelet coefficients are significant. Clearly an appropriate choice of the threshold value λ is fundamental to the effectiveness of the estimation procedure. Too large threshold might "cut off" important parts of the true function underlying the data while too small a threshold retains noise in the selective reconstruction. As described in Section 5.1, there are three commonly used thresholding functions. It has been shown that hard thresholding results in larger variance in the function estimate while soft thresholding has large bias. To comprise the trade-off between bias and variance, Bruce and Gao [27] suggested a firm thresholding that combines the hard and soft thresholding.

In the rest of the section, we discuss more literatures on the choice of thresholding for nonlinear regression. Donoho and Johnstone [58] proposed the universal threshold $\lambda_{un} = \sigma\sqrt{2\log n}/\sqrt{n}$ where σ is the noise level and can be estimated from the data. They also showed that for both hard and soft thresholding the resulting nonlinear wavelet estimator is asymptotically near-minimax in terms of L^2 risk and it outperforms any linear estimator for inhomogeneous functions. They [59] also proposed an adaptive SureShrink thresholding rule based on minimizing Stein's unbiased risk estimate. Papers [123; 86] investigated using cross-validation approaches for the choice of threshold. Some researchers [2; 128] developed the approaches of thresholding by hypothesis testing the coefficients for a significant deviation from zero. Donoho et al. [61] proposed level-dependent thresholding where different thresholds are used on different levels. Some researchers [30; 76] proposed block thresholding where coefficients are thresholded in blocks rather than individually. Both modifications imply better asymptotic properties of the resulting wavelet estimators. Various Bayesian approaches for thresholding and nonlinear shrinkage has also been proposed [161; 4; 3; 159]. In the Bayesian approach, a prior distribution is imposed on wavelet coefficient and then the function is estimated by applying a suitable Bayesian rule to the resulting posterior distribution of the wavelet coefficients. Garofalakis and Gibbons [70] introduced a probabilistic thresholding scheme that deterministically retains the most important coefficients while randomly rounding the other coefficients either up to a larger value or down to zero. The randomized rounding enables unbiased and error-guaranteed Reconstruction of individual data values. Interested readers may refer to [162] for comprehensive reviews of Bayesian approaches for thresholding. More discussion on nonlinear regression can be found in [10].

6.4 Distributed Data Mining

Over the years, data set sizes have grown rapidly with the advances in technology, the ever-increasing computing power and computer storage capacity, the permeation of Internet into daily life and the increasingly automated business, manufacturing and scientific processes. Moreover, many of these data sets are, in nature, geographically distributed across multiple sites. To mine such large and distributed data sets, it is important to investigate efficient distributed algorithms to reduce the communication overhead, central storage requirements, and computation times. With the high scalability of the distributed systems and the easy partition and distribution of a centralized dataset, distribute clustering algorithms can also bring the resources of multiple machines to bear on a given problem as the data size scale-up. In a distributed environment, data sites may be *homogeneous*, i.e., different sites containing data for exactly the same set of features, or *heterogeneous*, i.e., different sites storing data for different set of features, possibly with some common features among sites. The orthogonal property of wavelet basis could play an important role in distributed data mining since the orthogonality guarantees correct and independent local analysis that can be used as a building-block for a global model. In addition, the compact support property of wavelets could be used to design parallel algorithms since the compact support guarantees the localization of wavelet and processing a region of data with wavelet does not affect the the data out of this region.

Kargupta et al.[92; 81] introduced the idea of performing distributed data analysis using wavelet-based *Collective Data Mining*(CDM) from heterogeneous sites. The main steps for the approach can be summarized as follows:

- choose an orthonormal representation that is appropriate for the type of data model to be constructed,
- generate approximate orthonormal basis coefficients at each local site,
- if necessary, move an approximately chosen sample of the datasets from each site to a single site and generate the approximate basis coefficients corresponding to non-linear cross terms,
- combine the local models, transform the model into the user described canonical representation and output the model.

The foundation of CDM is based on the fact that any function can be represented in a distributed fashion using an appropriate basis. If we use wavelet basis, The orthogonality guarantees correct and independent local analysis that can be used as a building-block for a global model. Hershberger et al. [81] presented applications of wavelet-based CDM methodology to multivariate regression and linear discriminant analysis. Experiments have shown that the results produced by CDM are comparable to those obtained with centralized methods and the communication cost was shown to be directly proportional to the number of terms in the function and independent of the sample size.

6.5 Similarity Search/Indexing

The problem of similarity search in data mining is: given a pattern of interest, try to find similar patterns in the data set based on some similarity measures. This task is most commonly used for time series, image and text data sets. For time series, for example, given the Xerox stock prices over last 7 days and wish to find the stocks that have similar behaviors. For image, given a sample image and wish to find similar images in a collection of image database. For text, given some keywords, wish to find relevant documents. More formally, A dataset is a set denoted $DB = \{X_1, X_2, \ldots, X_i, \ldots, X_N\}$, where $X_i = [x_0^i, x_1^i, \ldots, x_n^i]$ and a given pattern is a sequence of data points $Q = [q_0, q_1, \ldots, q_n]$. Given a pattern Q, the result set R

from the data set is $R = \{X_{i_1}, X_{i_2}, \ldots, X_{i_j}, \ldots, X_{i_m}\}$, where $\{i_1, i_2, \cdots, i_m\} \subseteq \{1, \cdots, N\}$, such that $D(X_{i_j}, Q) < d$. If we use Euclidean distance between X and Y as the distance function $D(X,Y)$, then,

$$D(X,Y) = \sqrt{\sum_j |x_j - y_j|^2}$$

which is the aggregation of the point to point distance of two patterns. Wavelets could be applied into similarity search in several different ways. First, wavelets could transform the original data into the wavelet domain as described in Section 5.2 and we may also only keep selective wavelet coefficients to achieve dimensionality reduction as in Section 5.3. The similarity search are then conducted in the transformed domain and could be more efficient. Although the idea here is similar to that reviewed in Section 5.2 and Section 5.3: both involves transforming the original data into wavelet domain and may also selecting some wavelet coefficients. However, it should be noted that here for the data set: to project the n-dimensional space into a k-dimensional space using wavelets, the same k-wavelet coefficients should be stored for objects in the data set. Obviously, this is not optimal for all objects. To find the k optimal coefficients for the data set, we need to compute the average energy for each coefficient. Second, wavelet transforms could be used to extract compact feature vectors and define new similarity measure to facilitate search. Third, wavelet transforms are able to support similarity search at different scales. The similarity measure could then be defined in an adaptive and interactive way.

Wavelets have been extensively used in similarity search in time series [83; 172; 131; 132]. Excellent overview of wavelet methods in time series analysis can be found in [44; 121; 122]. Chan and Fu [131] proposed efficient time series matching strategy by wavelets. Haar transform wavelet transform is first applied and the first few coefficients of the transformed sequences are indexed in an R-Tree for similarity search. The method provides efficient for range and nearest neighborhood queries. Huhtala et al. [83] also used wavelets to extract features for mining similarities in aligned time series. Wu et al.[172] presented a comprehensive comparison between DFT and DWT in time series matching. The experimental results show that although DWT does not reduce relative matching error and does not increase query precision in similarity search, DWT based techniques have several advantage such as DWT has multi-resolution property and DWT has complexity of $O(N)$ while DFT has complexity of $O(N \log N)$. Wavelet transform gives time-frequency localization of the signal and hence most of the energy of the signal can be represented by only a few DWT coefficients. Struzik and Siebes [153; 154] presented new similarity measures based on the special presentations derived from Haar wavelet transform. Instead of keeping selective wavelet coefficients, the special representations keep only the sign of the wavelet coefficients (sign representation) or keep the difference of the logarithms (DOL) of the values of the wavelet coefficient at highest scale and the working scale (DOL representation). The special representations are able to give step-wise comparisons of correlations and it was shown that the similarity measure based on such representations closely corresponds to the subjective feeling of similarity between time series.

Wavelets also have widespread applications in content-based similarity search in image/audio databases. Jacobs et al.[85] presented a method of using *image querying metric* for fast and efficient content-based image querying. The image querying metric is computed on the *wavelet signatures* which are obtained by truncated and quantized wavelet decomposition. In essential, the image querying metric compares how many wavelet significant wavelet coefficients the query has in common with the potential targets. Natsev et al. [125] proposed *WALRUS* (WAveLet-based Retrieval of User-specified Scenes) algorithm for similarity retrieval in image diastases. WALRUS first uses dynamic programming to compute wavelet signatures for sliding windows of varying size, then clusters the signatures in wavelet space and finally the similarity measure between a pair of images is calculated to be the fraction of the area the two images covered by matching signatures. Ardizzoni et al. [13] described *Windsurf* (Wavelet-Based Indexing of Images Using Region Fragmentation), a new approach for image retrieval. Windsurf uses Haar wavelet transform to extract color and texture features and applies clustering techniques to partition the image into regions. Similarity is then computed as the Bhattcharyya metric [31] between matching regions. Brambilla [24] defined an effective strategy which exploits multi-resolution wavelet transform to effectively describe image content and is capable of interactive learning of the similarity measure. Wang et al. [167; 84] described *WBIIS* (Wavelet-Based Image Indexing and Searching), a new image indexing and retrieval algorithm with partial sketch image searching capability for large image databases. WBIIS applies Daubechies-8 wavelets for each color component and low frequency wavelet coefficients and their variance are stored as feature vectors. Wang, Wiederhold and Firschein [166] described *WIPETM* (Wavelet Image Pornography Elimination) for image retrieval. WIPETM uses Daubechies-3 wavelets, normalized central moments and color histograms to provide feature vector for similarity matching. Subramanya and Youssef [155] presented a scalable content-based image indexing and retrieval system based on vector coefficients of color images where highly decorrelated wavelet coefficient planes are used to acquire a search efficient feature space. Mandal et al. [112] proposed fast wavelet histogram techniques for image indexing. There are also lots of applications of wavelets in audio/music information processing such as [103; 56; 101; 156]. In fact, IEEE Transactions on Signal Processing has two special issues on wavelets, in Dec. 1993 and Jan. 1998 respectively. Interested readers could refer to these issues for more details on wavelets for indexing and retrieval in signal processing.

6.6 Approximate Query Processing

Query processing is a general task in data mining and similarity search discussed in Section 6.5 is one of the specific form of query processing. In this section, we will describe wavelet applications in approximate query processing which is another area within query processing. Approximate query processing has recently emerged as a viable solution for large-scale decision support. Due to the exploratory nature of many decision support applications, there are a number of scenarios where an exact answer may not be required and a user may in fact prefer a fast approximate answer. Wavelet-based techniques can be applied as a data reduction mechanism to obtain *wavelet synopses* of the data on which the approximate query could then operate. The wavelet synopses are compact sets of wavelet coefficients obtained by the wavelet decomposition. Note that some of wavelet methods described here might overlap with those described in Section 5.3. The wavelet synopses reduce large amount of data to compact sets and hence could provide fast and reasonably approximate answers to queries.

Matias, Vitter and Wang [113; 114] presented a wavelet-based technique to build histograms on the underlying data distributions for selectivity estimation and Vitter et al. [164; 88] also proposed wavelet-based techniques for the approximation of range-sum queries over OLAP data cubes. Generally, the central idea is to apply multidimensional wavelet decomposition on the input data collection (attribute columns or OLAP cube) to obtain a compact data synopsis by keeping a selective small collection of wavelet coefficients. Experiments in [113] showed that wavelet-based histograms improve the accuracy substantially over random sampling and results from [164] clearly demonstrated that wavelets can be very effective in handling aggregates over high-dimensional OLAP cubes while avoiding the high construction costs and storage overheads. Chakrabarti et al. [36] extended previous work on wavelet techniques in approximate query answering by demonstrating that wavelets could be used as a generic and effective tool for decision support applications. The generic approach consists of three steps: the wavelet-coefficient synopses are first computed and then using novel query processing algorithms SQL operators such as select, project and join can be executed entirely in the wavelet-coefficient domain. Finally the results is mapped from the wavelet domain to relational tuples(*Rendering*). Experimental results verify the effectiveness and efficiency. Gilbert et al. [71] presented techniques for computing small space representations of massive data streams by keeping a small number of wavelet coefficients and using the representations for approximate aggregate queries. Garofalakis and Gibbons [70] introduced probabilistic wavelet synopses that provably enabled unbiased data reconstruction with guarantees on the accuracy of individual approximate answers. The probabilistic technique is based on probabilistic thresholding scheme to assign each coefficient a probability of being retained instead of deterministic thresholding.

6.7 Visualization

Visualization is one of the description tasks (exploratory data analysis) of data mining and it allows the user to gain an understanding of the data. Visualization works because it exploits the broader information bandwidth of graphics as opposed to text or numbers. However, for large dataset it is often not possible to even perform simple visualization task. The multiscale wavelet transform facilitates progressive access to data with the viewing of the most important features first.

Miller et al. [118] presented a novel approach to visualize and explore unstructured text based on wavelet. The underlying technology applies wavelet transforms to a custom digital signal constructed from words within a document. The resultant multiresolution wavelet energy is used to analyze the characteristics of the narrative flow in the frequency domain. Wong and Bergeron [170] discussed with the authenticity issues of the data decomposition, particularly for data visualization. A total of six datasets are used to clarify the approximation characteristics of compactly supported orthogonal wavelets. It also presents an error tracking mechanism, which uses the available wavelet resources to measure the quality of the wavelet approximations. Roerdink and Westenberg [137] considered multiresolution visualization of large volume data sets based on wavelets. Starting from a wavelet decomposition of the data, a low resolution image is computed; this approximation can be successively refined. Du and Moorhead [62] presented a technique which used a wavelet transform and MPI(Message Passing Interface) to realize a distributed visualization system. The wavelet transform has proved to be a useful tool in data decomposition and progressive transmission.

6.8 Neural Network

Neural networks are of particular interest because they offer a means of efficiently modeling large and complex problems and they can be applied to many data mining tasks such as classification, clustering and regression. Roughly speaking, a neural network is a set of connected input/hidden/output units where each connection has an associated weight and each unit has an associated activated function. Usually neural network methods contain a learning phase and a working phase. A learning phase is to adjust the weights and the structures of the network based on the training samples while the working phase is to execute various tasks on new instances. For more details on neural network, please refer to [80; 64; 90].

The idea of combining neural networks with multiscale wavelet decomposition has been proposed by a number of authors [42; 98; 43; 54; 97; 49; 95; 96; 171]. These approaches either use wavelets as the neuron's activation functions [98; 38](usually call these as wavelet neural network), or in a pre-processing phasing by the extraction of features from time series data [42; 54; 171]. The properties of wavelet transforms emerging from a multi-scale decomposition of signals allow the study of both stationary and non-stationary signals. On the other hand the neural network performs a non-linear analysis as well linear dependencies due to different possible structures and activation functions. Hence combining wavelets and neural network would give us more power on data analysis. A wavelet neural network, using the wavelets as activation functions and combining the mathematically rigorous, multi-resolution character of wavelets with the adaptive learning of artificial neural networks, has the capability of approximating any continuous nonlinear mapping to any high resolution. Learning with wavelet neural network is efficient, and is explicitly based on the local or global error of approximation. A simple wavelet neural network displays a much higher level of generalization and shorter computing time as compared to three-layer feed forward neural network [173]. Roverso [139] proposed an approach for multivariate temporal classification by combining wavelet and recurrent neural network. Kreinovich et al. [99] showed that wavelet neural networks are asymptotically optimal approximators for functions of one variable in the sense that it require to store the smallest possible number of bits that is necessary to reconstruct a function with a given precision. Bakshi et al. [18] described the advantages of wavelet neural network learning over other artificial neural learning techniques and discussed the relationship between wavelet neural network and other rule-extraction techniques such as decision trees. It also shows that wavelets may provide a unifying framework for various supervised learning techniques.

WSOM is a feedforward neural network that estimates optimized wavelet bases for the discrete wavelet transform on the basis of the distribution of the input data [32]. Sheng and Chou [105] reported the application of using wavelet transform and self-organizing map to mine air pollutant data.

6.9 Principal/Independent Component Analysis

A widely used technique for data mining is based on diagonalizing the correlation tensor of the data-set, keeping a small number of coherent structures (eigenvectors) based on principal components analysis (PCA) [19]. This approach tends to be global in character. Principal component analysis (PCA) has been adopted for many different tasks. Wavelet analysis and PCA can be combined to obtain proper accounting of global contributions to signal energy without loss of information on key local features. In addition, the multi-resolution property of wavelets could help to find the principal component at multiple scales.

Bakshi [16] used multiscale PCA(MSPCA) for process monitoring. Multiscale PCA combines the ability of PAC to decorrelate the variables by extracting a linear relationship with that of wavelet analysis to extract deterministic features and approximately decorrelate autocorrelated measurements. MSPCA computes the PCA of the wavelet coefficients at each scale, followed by combining the results at relevant scales. Due to its multiscale nature, MSPCA is approximate for modeling of data containing contributions from events whose behavior changes over time and frequency. Process monitoring by MSPCA involves combining only those scales where significant events are detected, and is equivalent to adaptively filtering the scores and residuals and adjusting limits for easiest detection of deterministic changes in the measurements. Bakshi [17] presented an overview of multiscale data analysis and empirical modeling methods based on wavelet analysis. Feng et al. [65] proposed an approach of applying Principal Component Analysis (PCA) on the wavelet subband as described in Section 5.2. Wavelet analysis could also be combined with Independent Component Analysis (ICA). The goal of ICA is to recover independent sources given only sensor observations that are unknown linear mixtures of the unobserved independent source signals. Briefly, ICA attempts to estimate the coefficients of an unknown mixture of n signal sources under the hypotheses that the sources are statistically independent, the medium of transmission is deterministic, and crucially, the mixture coefficients are constant with respect to time. One then solves for the sources from the observations by *inverting* the mixture matrix. In contrast to correlation-based transformations such as Principal Component Analysis (PCA), ICA not only decorrelates the signals (2nd-order statistics) but also reduces higher-order statistical dependencies, attempting to make the signals as independent as possible. In other words, ICA is a way of finding a linear non-orthogonal co-ordinate system in any multivariate data. The directions of the axes of this co-ordinate system are determined by both the second and higher order statistics of the original data. The goal is to perform a linear transform which makes the resulting variables as statistically independent from each other as possible. More details about the ICA algorithms can be found in [21; 48; 20; 89]. A fundamental weakness of existing ICA algorithms, namely that the mixture matrix is assumed to be essentially constant. This is unsatisfactory when moving sources are involved. Wavelet transforms can be utilized to this problem by using time-frequency characteristics of the mixture matrix in the source identification. Moreover, ICA algorithms could also make use of the multiscale representation of wavelet transforms.

7. SOME OTHER APPLICATIONS

There are some other wavelet applications that are related to data mining.

Web Log Mining: Wavelets offer powerful techniques for mathematically representing web requests at multiple time scales and a compact and concise representation of the requests using wavelet coefficients. Zhai et al. [175] proposed to use wavelet-based techniques to analyze the workload collected from busy web servers. It aims at finding the temporal characteristics of the web server weblog which contains workload information and predicting the trend it evolves.

Traffic Monitoring: The wavelet transform significantly reduces the temporal dependence and simple models which are insufficient in the time domain may be quite accurate in the wavelet domain. Hence wavelets provide an efficient way to modeling network traffic. Riedi et al. [136] developed a new multiscale modeling framework for characterizing positive-valued data with long-range-dependent correlations. Using the Haar wavelet transform and a special multiplicative structure on the wavelet and scaling coefficients to ensure positive results. Ma and Ji [106; 107; 108] presented the work on modeling on modeling temporal correlation (the second-order statistics) of heterogeneous traffic, and modeling non-Gaussian (high-order statistics) and periodic traffic in wavelet domain.

Change Detection: The good time-frequency localization of wavelets provides a natural motivation for their use in change point detection problems. The main goal of change detection is estimation of the number, locations and sizes of function's abrupt changes such as sharp spikes or jumps. Change-point models are used in a wide set of practical problems in quality control, medicine, economics and physical sciences [1]. The general idea of using wavelet for detecting abrupt changes is based on the connection between the function's local regularity properties at a certain point and the rate of decay of the wavelet coefficients located near this point across increasing resolution level [111]. Local regularities are identified by unusual behavior in the wavelet coefficients at high-resolution levels at the corresponding location [168]. Bailey et al. [15] used wavelet to detect signal in underwater sound. Donoho et al. [57] discussed the application of wavelets for density estimation.

8. CONCLUSION

This paper provides an application-oriented overview of the mathematical foundations of wavelet theory and gives a comprehensive survey of wavelet applications in data mining The object of this paper is to increase familiarity with basic wavelet applications in data mining and to provide reference sources and examples where the wavelets may be usefully applied to researchers working in data analysis. Wavelet techniques have a lot of advantages and there already exists numerous successful applications in data mining. It goes without saying that wavelet approaches will be of growing importance in data mining.

It should also be mentioned that most of current works on wavelet applications in data mining are based orthonormal wavelet basis. However, we argue that orthonormal basis may not be the best representation for noisy data even though the vanishing moments can help them achieve denoising and dimensionality reduction purpose. Intuitively, orthogonality is the most economical representation. In other words, in each direction, it contains equally important information. Therefore, it is usually likely that thresholding wavelet coefficients remove useful information when they try to remove the noise or redundant information (noise can also be regarded as one kind of redundant information). To represent redundant information, it might be good to use redundant wavelet representation – wavelet frames. Except orthogonality, wavelet frames preserve all other properties that an orthonormal wavelet basis owns, such as vanishing moment, compact support, multiresolution. The redundancy of a wavelet frame means that the frame functions are not independent anymore. For example, vectors $[0, 1]$ and $[1, 0]$ is an orthonormal basis of a plane R^2, while vectors $[1/2, 1/2]$, $[-1/2, 1/2]$, and $[0, -1]$ are not independent, and consist a frame for R^2. So when data contain noise, frames may provide some specific directions to record the noise. Our work will be the establishment of criteria to recognize the direction of noise or redundant information.

Wavelets could also potentially enable many other new researches and applications such as conventional database compression, multiresolution data analysis and fast approximate data mining etc. Finally we eagerly await many future developments and applications of wavelet approaches in data mining.

9. REFERENCES

[1] F. Abramovich, T. Bailey, and T. Sapatinas. Wavelet analysis and its statistical applications. *JRSSD*, (48):1–30, 2000.

[2] F. Abramovich and Y. Benjamini. Thresholding of wavelet coecients as multiple hypotheses testing procedure. In

A. Antoniadis and G. Oppenheim, editors, *Wavelets and Statistics, Lecture Notes in Statistics 103*, pages 5–14. Springer-Verlag, New York, 1995.

[3] F. Abramovich and T. Sapatinas. Bayesian approach to wavelet decomposition and shrinkage. In P. Muller and B. Vidakovic, editors, *Lecture Notes in Statistics*. Springer-Verlag, New York, 1999.

[4] F. Abramovich, T. Sapatinas, and B. Silverman. Wavelet thresholding via a Bayesian approach. *Journal of the Royal Statistical Society, Series B*, (58), 1997.

[5] C. C. Aggarwal, J. L. Wolf, P. S. Yu, C. Procopiuc, and J. S. Park. Fast algorithms for projected clustering. pages 61–72, 1999.

[6] C. C. Aggarwal and P. S. Yu. Outlier detection for high dimensional data. In *SIGMOD Conference*, 2001.

[7] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan. Automatic subspace clustering for high dimensional data for data mining applications. In *SIGMOD-98*, 1998.

[8] H. Akaike. Information theory and an extension of the maximum likelihood principle. In *In Proc. 2nd Int. Symp. Info. Theory*, pages 267–281, 1973.

[9] A. Aldroubi and M. Unser, editors. *Wavelets in Medicine and Biology*. CRC Press, Boca Raton, 1996.

[10] A. Antoniadis. Wavelets in statistics: a review. *J. It. Statist. Soc.*, 1999.

[11] A. Antoniadis, G. Grégoire, and I. W. McKeague. Wavelet methods for curve estimation. *Journal of the American Statistical Association*, 89(428):1340–1353, 1994.

[12] A. Antoniadis and G. Oppenhiem, editors. *Wavelets and Statistics, Lecture Notes in Statistics*. Springer-Verlag, 1995.

[13] S. Ardizzoni, I. Bartolini, and M. Patella. Windsurf: Region-based image retrieval using wavelets. In *DEXA Workshop*, pages 167–173, 1999.

[14] A. Arning, R. Agrawal, and P. Raghavan. A linear method for deviation detection in large databases. In *Knowledge Discovery and Data Mining*, pages 164–169, 1996.

[15] T. C. Bailey, T. Sapatinas, K. J. Powell, and W. J. Krzanowski. Signal detection in underwater sounds using wavelets. *Journal of the American Statistical Association*, 93(441):73–83, 1998.

[16] B. Bakshi. Multiscale pca with application to multivariate statistical process monitoring. *AIChE Journal*, 44(7):1596–1610, 1998.

[17] B. Bakshi. Multiscale analysis and modeling using wavelets. *Journal of Chemometrics*, (13):415–434, 1999.

[18] B. R. Bakshi, A. Koulouris, and G. Stephanopoulos. Learning at multiple resolutions: Wavelets as basis functions in artificial neural networks and inductive decision trees. In R. Motard and B. Joseph, editors, *Wavelet Applications in Chemical Engineering*. Kluwer Inc., Boston, 1994.

[19] D. Ballard. *An introduction to natural computation*. MIT Press, 1997.

[20] A. Bell and T. Sejnowski. Fast blind separation based on information theory. In *Proc. Intern. Symp. on Nonlinear Theory and Applications*, Las Vegas, 1995.

[21] A. J. Bell and T. J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159, 1995.

[22] M. Berger and I. Rigoutsos. An algorithm for point clustering and grid generation. *IEEE Trans. on Systems, Man and Cybernetics*, 21(5):1278–1286, 1991.

[23] M. Blume and D. Ballard. Image annotation based on learning vector quantization and localized haar wavelet transform features. In *Proc. SPIE 3077*, pages 181–190, 1997.

[24] C. Brambilla, A. D. Ventura, I. Gagliardi, and R. Schettini. Multiresolution wavelet transform and supervised learning for content-based image retrieval. In *Proceedings of the IEEE International Conference on Multimedia Computing and Systems*, volume I, 1999.

[25] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander. LOF: identifying density-based local outliers. In *Proceedings of ACMSIGMOD Conference*, pages 93–104, 2000.

[26] M. Brito, E. Chavez, A. Quiroz, and J. Yukich. Connectivity of the mutual K-Nearest-Neighbor graph for clustering and outlier detection. *Statistics and Probability Letters*, 35:33–42, 1997.

[27] A. Bruce and H.-Y. Gao. Waveshrink with firm shrinkage. *Statistica Sinica*, (4):855–874, 1996.

[28] I. Bruha and A. F. Famili. Postprocessing in machine learning and data mining. *SIGKDD Exlporations*, 2000.

[29] R. W. Buccigrossi and E. P. Simoncelli. Image compression via joint statistical characterization in the wavelet domain. In *Proceedings ICASSP-97 (IEEE International Conference on Acoustics, Speech and Signal Processing)*, number 414, Munich, Germany, 1997.

[30] T. Cai. Adaptive wavelet estimation: a block thresholding and oracle inequality approach. Technical Report 98-07, Department of Statistics,Purdue University, 1998.

[31] J. P. Campbell. Speaker recognition: A tutorial. In *Proceedings of the IEEE*, volume 85, pages 1437–1461, Sept. 1997.

[32] G. Carpenter. Wsom: building adaptive wavelets with self-organizing maps. In *Proc. of 1998 IEEE International Joint Conference on Neural Networks*, volume 1, pages 763–767, 1998.

[33] V. Castelli and I. Kontoyiannis. Wavelet-based classification: Theoretical analysis. Technical Report RC-20475, IBM Watson Research Center, 1996.

[34] V. Castelli and I. Kontoyiannis. An efficient recursive partitioning algorithm for classification, using wavelets. Technical Report RC-21039, IBM Watson Research Center, 1997.

[35] V. Castelli, C. Li, J. Turek, and I. Kontoyiannis. Progressive classification in the compressed domain for large eos satellite databases, April 1996.

[36] K. Chakrabarti, M. Garofalakis, R. Rastogi, and K. Shim. Approximate query processing using wavelets. *VLDB Journal: Very Large Data Bases*, 10(2-3):199–223, 2001.

[37] A. Chambolle, R. DeVore, N. Lee, and B. Lucier. Nonlinear wavelet image processing: Variational problems, compression, and noise removal through wavelet shrinkage. *IEEE Tran. Image Proc.*, 7(3):319–333, 1998.

[38] P.-R. Chang and B.-F. Yeh. Nonlinear communication channel equalization using wavelet neural networks. In *Proc. of 1994 IEEE International Conference Joint on Neural Networks*, volume 6, pages 3605–3610, 1994.

[39] S. Chang, B. Yu, and M. Vetterli. Spatially adaptive wavelet thresholding with context modeling for image denoising. In *ICIP*, volume 1, pages 535–539, 1998.

[40] T. Chang and C. Kuo. Texture analysis and classification with tree-structured wavelet transform. *IEEE Trans. on Image Processing*, 2(4):429–441, 1993.

[41] P. Cheeseman, J. Kelly, and M. Self. AutoClass: A bayesian classification system. In *ICML'88*, 1988.

[42] B. Chen, X.Z.Wang, S. Yang, and C. McGreavy. Application of wavelets and neural networks to diagnostic system development,1,feature extraction. *Computers and Chemical Engineering*, (23):899–906, 1999.

[43] B. Chen, X.Z.Wang, S. Yang, and C. McGreavy. Application of wavelets and neural networks to diagnostic system development,2,an integrated framework and its application. *Computers and Chemical Engineering*, (23):945–954, 1999.

[44] C. Chiann and P. A. Morettin. A wavelet analysis for time series. *Journal of Nonparametric Statistics*, 10(1):1–46, 1999.

[45] C. Chrysafis and A. Ortega. Line based, reduced memory, wavelet image compression. In *Data Compression Conference*, pages 398–407, 1998.

[46] C. K. Chui. *An Introduction to Wavelets*. Academic Press, Boston, 1992.

[47] C. K. Chui and J. Lian. A study of orthonormal multiwavelets. *Applied Numerical Mathematics: Transactions of IMACS*, 20(3):273–298, 1996.

[48] P. Comon. Independent component analysis - a new concept? *Signal Processing*, (36):287–314, 1994.

[49] P. Cristea, R. Tuduce, and A. Cristea. Time series prediction with wavelet neural networks. In *Proc. of the 5th Seminar on Neural Network Applications in Electrical Engineering*, pages 5–10, 2000.

[50] R. F. Cromp and W. J. Campbell. Data mining of multidimensional remotely sensed images. In *Proc. 2nd International Conference of Information and Knowledge Management,*, Arlington, VA, Nov 1993.

[51] I. Daubechies. Orthonormal bases of compactly support wavelets. *Comm. Pure Applied Mathematics*, 41:909–996, 1988.

[52] I. Daubechies. *Ten Lectures on Wavelets*. Capital City Press, Montpelier, Vermont, 1992.

[53] I. Daubechies, B. Han, A. Ron, and Z. Shen. Framelets: Mrabased constructions of wavelet frames, 2000. preprint.

[54] C. J. Deschenes and J. P. Noonan. A fuzzy kohonen network for the classification of transients using the wavelet transform for feature extraction. *Information Sciences*, (87):247–266, 1995.

[55] R. A. DeVore, B. Jawerth, and B. J. Lucier. image compression through wavelet transform coding. *IEEE Transactions on Information Theory*, 38(2):719–746, 1992.

[56] P. Q. Dinh, C. Dorai, and S. Venkatesh. Video genre categorization using audio wavelet coefficients. In *ACCV 2002*, 2002.

[57] D. Donoho, I. Johnstone, G. Kerkyacharian, and D. Picard. Density estimation by wavelet thresholding. *Ann. Statist.*, (24):508–539, 1996.

[58] D. L. Donoho and I. M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994.

[59] D. L. Donoho and I. M. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, 90(432):1200–1224, 1995.

[60] D. L. Donoho and I. M. Johnstone. Minimax estimation via wavelet shrinkage. *Annals of Statistics*, 26(3):879–921, 1998.

[61] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard. Wavelet shrinkage: Asymptopia? *J. R. Statist. Soc. B.*, 57(2):301–337, 1995.

[62] X. S. Du and R. J. Moorhead. Multiresolutional visualization of evolving distributed simulations using wavelets and mpi. In *SPIE EI'97*, 1997.

[63] G. Fan and X. Xia. Wavelet-based statistical image processing using hidden markov tree model. In *Proc. 34th Annual Conference on Information Sciences and Systems, Princeton, NJ, USA*, 2000.

[64] L. Fausett. *Fundamentals of Neural Networks*. Prentice Hall, 1994.

[65] G. C. Feng, P. C. Yuen, and D. Q. Dai. Human face recognition using PCA on wavelet subband. *SPIE Journal of Electronic Imaging*, 9(2), 2000.

[66] D. H. Fisher. Iterative optimization and simplification of hierarchical clusterings. Technical Report CS-95-01, Vanderbilt U., Dept. of Comp. Sci., 1995.

[67] P. Flandrin. Wavelet analysis and synthesis of fractional Brownian motion. *IEEE Transactions on Information Theory*, 38(2):910–917, 1992.

[68] V. Ganti, J. Gehrke, and R. Ramakrishnan. CACTUS - clustering categorical data using summaries. In *Knowledge Discovery and Data Mining*, pages 73–83, 1999.

[69] H.-Y. Gao. Threshold selection in WaveShrink, 1997. theory for matlab wavelet toolbox on denoising.

[70] M. Garofalakis and P. B. Gibbons. Wavelet synopses with erro guarantee. In *Proceedings of 2002 ACM SIGMOD*, Madison, Wisconsin, USA, June 2002. ACM Press.

[71] A. C. Gilbert, Y. Kotidis, S. Muthukrishnan, and M. Strauss. Surfing wavelets on streams: One-pass summaries for approximate aggregate queries. In *The VLDB Journal*, pages 79–88, 2001.

[72] S. Guha, R. Rastogi, and K. Shim. CURE: an efficient clustering algorithm for large databases. In *Proceedings of ACM SIGMOD*, pages 73–84, 1998.

[73] S. Guha, R. Rastogi, and K. Shim. ROCK: A robust clustering algorithm for categorical attributes. *Information Systems*, 25(5):345–366, 2000.

[74] D. Gunopulos. Tutorial slides: Dimensionality reduction techniques. In DIMACS Summer School on Data Mining, August 2001.

[75] A. Haar. Zur theorie der orthogonalen funktionensysteme. *Mathematics Annal.*, 69:331–371, 1910.

[76] P. Hall, G. Kerkyacharian, and D. Picard. Block threshold rules for curve estimation using kernel and wavelet methods. *Ann. Statist.*, (26):922–942, 1998.

[77] J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers, 2000.

[78] J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers, 2000.

[79] D. Hand, H. Mannila, and P. Smyth. *Principles of Data Mining*. The MIT Press, 2001.

[80] S. Haykin. *Neural Networks*. Prentice Hall, 1999.

[81] D. E. Hershberger and H. Kargupta. Distributed multivariate regression using wavelet-based collective data mining. *Journal of Parallel and Distributed Computing*, 61(3):372–400, 2001.

[82] P. Hornby, F. Boschetti, and F. Horowitz. Analysis of potential field data in the wavelet domain. In *Proceedings of the 59th EAGE Conference*, May 1997.

[83] Y. Huhtala, J. Karkkainen, and H. Toivonen. Mining for similarities in aligned time series using wavelets. In *Data Mining and Knowledge Discovery: Theory, Tools, and Technology. SPIE Proc.*, 1999.

[84] O. F. J. Z. Wang, G. Wiederhold and S. X. Wei. Wavelet-based image indexing techniques with partial sketch retrieval capability. *IEEE Advances in Digital Libraries*, 1(4):311–328, May 1997.

[85] C. E. Jacobs, A. Finkelstein, and D. H. Salesin. Fast multiresolution image querying. *Computer Graphics*, 29(Annual Conference Series):277–286, 1995.

[86] M. Jansen, M. Malfait, and A. Bultheel. Generalized cross validation for wavelet thresholding, 1995. preprint, Dec. 1995.

[87] W. Jin, A. K. H. Tung, and J. Han. Mining top-n local outliers in large databases. In *Knowledge Discovery and Data Mining*, pages 293–298, 2001.

[88] J.S.Vitter, M. Wang, and B. Iyer. Data cube approximation and histograms via wavelets. In *Proc. of the 7th Intl. Conf. On Infomration and Knowledge Management*, 1998.

[89] C. Jutten, J. Herault, P. Comon, and E. Sorouchiary. Blind separation of sources, Parts I, II and III. *Signal Processing*, (24):1–29, 1991.

[90] G. K. *An Introduction to Neural Networks*. UCL Press, 1997.

[91] L. Kaplan and C. Kuo. Fractal estimation from noisy data via discrete fractional Gaussian noise (DFGN) and the Harr basis. *IEEE Transactions on Information Theory*, 41(12):3554–3562, 1993.

[92] H. Kargupta and B. Park. The collective data mining: A technology for ubiquitous data analysis from distributed heterogeneous sites, 1998. Submitted to IEEE Computer Special Issue on Data Mining.

[93] D. Keim and M. Heczko. Wavelets and their applications in databases. Tutorial Notes of ICDE 2001, 2001.

[94] E. M. Knorr and R. T. Ng. Finding intensional knowledge of distance-based outliers. In *The VLDB Journal*, pages 211–222, 1999.

[95] K. Kobayashi and T. Torioka. A wavelet neural network for function approximation and network optimization. In *Proceedings of ANNIE'94, AMSE Press.*, pages 505–510, 1994.

[96] K. Kobayashi and T. Torioka. Designing wavelet networks using genetic algorithms. In *Proceedings of EUFIT'97*, volume 1, 1997.

[97] P. Kostka, E. Tkacz, Z. Nawrat, and Z. Malota. An application of wavelet neural networks (wnn) for heart valve prostheses characteristic. In *Proc. of the 22nd Annual International Conference of IEEE, Engineering in Medicine and Biology Society*, volume 4, pages 2463–2465, 2000.

[98] A. Koulouris, B. R. Bakshi, and G. Stephanopoulos. Empirical learning through neural networks: The wave-net solution. *Intelligent Systems in Process Engineering*, pages 437–484, 1996.

[99] V. Kreinovich, O. Sirisaengtaksin, and S. Cabrera. Wavelet neural networks are optimal approximators for functions of one variable. Technical Report 29, University of Texas at El Paso/University of Houston, 1992.

[100] A. Laine and J. Fan. texture classification by wavelet packet signatures. Technical report, University of Florida, 1992.

[101] T. Lambrou, P. Kudumakis, R. Speller, M. Sandler, and A. Linney. Classification of audio signals using statistical features on time and wavelet tranform domains. In *Proc. Int. Conf. Acoustic, Speech, and Signal Processing (ICASSP-98)*, volume 6, pages 3621–3624, 1998.

[102] P. C. Lemarié and Y. Meyer. Ondelettes et bases hilbertiennes. *Rev. Mat. Ibero-Amer*, pages 1–18, 1986.

[103] G. Li and A. A. Khokhar. Content-based indexing and retrieval of audio data using wavelets. In *IEEE International Conference on Multimedia and Expo (II)*, pages 885–888, 2000.

[104] Q. Li, T. Li, and S. Zhu. Improving medical/biological data classification performance by wavelet pre-processing. In *ICDM 2002*, 2002.

[105] S.-T. Li and S.-W. Chou. Multi-resolution spatio-temporal data mining for the study of air pollutant regionalization. In *Proceedings of the 33rd Hawaii International Conference on System Sciences*, 2000.

[106] S. Ma and C. Ji. Modeling heterogeneous network traffic in wavelet domain: Part I-temporal correlation. Technical report, 1999.

[107] S. Ma and C. Ji. Modeling heterogeneous network traffic in wavelet domain: Part II-non-gaussian traffic. Technical report, IBM Waston Research Center, 1999.

[108] S. Ma and C. Ji. Modeling heterogeneous network traffic in wavelet domain. *IEEE/ACM Transactions on Networking*, 9(5), October 2001.

[109] S. Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):674–693, 1989.

[110] S. Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, San Diego, 1998.

[111] S. G. Mallat and W. L. Hwang. Singularity detection and processing with wavelets. *IEEE Transactions on Information Theory*, 38(2):617–643, 1992.

[112] M. K. Mandal, T. Aboulnasr, and S. Panchanathan. Fast wavelet histogram techniques for image indexing. *Computer Vision and Image Understanding: CVIU*, 75(1–2):99–110, 1999.

[113] Y. Matias, J. S. Vitter, and M. Wang. Wavelet-based histograms for selectivity estimation. In *ACM SIGMOD*, pages 448–459. ACM Press, 1998.

[114] Y. Matias, J. S. Vitter, and M. Wang. Dynamic maintenance of wavelet-based histograms. In *VLDB'00*. Morgan Kaufmann, 2000.

[115] P. Meerwald and A. Uhl. A survey of wavelet-domain watermarking algorithms. In *Proceedings of SPIE, Electronic Imaging, Security and Watermarking of Multimedia Contents III*, volume 4314, San Jose, CA, USA, jan 2001. SPIE.

[116] Y. Meyer. *Wavelets and Operators*. Cambridge University Press, 1992.

[117] Y. Meyer. *Wavlets–Algorithms and Application*. SIAM, 1993.

[118] N. E. Miller, P. C. Wong, M. Brewster, and H. Foote. TOPIC ISLANDS - A wavelet-based text visualization system. In D. Ebert, H. Hagen, and H. Rushmeier, editors, *IEEE Visualization '98*, pages 189–196, 1998.

[119] T. M. Mitchell. *Machine Learning*. The McGraw-Hill Companies,Inc., 1997.

[120] A. Mojsilovic and M. v. Popovic. Wavelet image extension for analysis and classification of infarcted myocardial tissue. *IEEE Transactions on Biomedical Engineering*, 44(9), September 1997.

[121] P. Morettin. Wavelets in statistics. (3):211–272, 1997.

[122] P. A. Morettin. From fourier to wavelet analysis of time series. In A. Prat, editor, *Proceedings in Computational Statistics*, pages 111–122, 1996.

[123] G. P. Nason. Wavelet shrinkage by cross-validation. *Journal of the Royal Statistical Society B*, 58:463–479, 1996.

[124] G. P. Nason and R. von Sachs. Wavelets in time series analysis. *Philosophical Transactions of the Royal Society of London A*, 357(1760):2511–2526, 1999.

[125] A. Natsev, R. Rastogi, and K. Shim. Walrus:a similarity retrieval algorithm for image databases. In *Proceedings of ACM SIGMOD International Conference on Management of Data*, pages 395–406. ACM Press, 1999.

[126] R. T. Ng and J. Han. Efficient and effective clustering methods for spatial data mining. In J. Bocca, M. Jarke, and C. Zaniolo, editors, *20th International Conference on Very Large Data Bases, September 12–15, 1994, Santiago, Chile proceedings*, pages 144–155, Los Altos, CA 94022, USA, 1994. Morgan Kaufmann Publishers.

[127] R. Ogden. *Essential Wavelets for Statistical Application and Data Analysis*. Birkhauser, Boston, 1997.

[128] R. T. Ogden and E. Parzen. Data dependent wavelet thresholding in nonparametric regression with change-point applications. *Computational Statistics & Data Analysis*, 22:53–70, 1996.

[129] D. Percival and A. T. Walden. *Wavelet Methods for Time Series Analysis*. Cambridge University Press, 2000.

[130] R. Polikar. The wavelet tutorial. Internet Resources: http://engineering.rowan.edu/ polikar/WAVELETS/WTtutorial.html.

[131] K. pong Chan and A. W.-C. Fu. Efficient time series matching by wavelets. In *ICDE*, pages 126–133, 1999.

[132] I. Popivanov and R. J. Miller. Similarity search over time series data using wavelets. In *ICDE 2002*, 2002.

[133] L. Prasad, S. S. Iyengar, and S. S. Ayengar. *Wavelet Analysis with Applications to Image Processing*. CRC Press, 1997.

[134] D. Pyle. *Data Preparation for Data Mining*. Morgan Kaufmann, 1999.

[135] S. Ramaswamy, R. Rastogi, and K. Shim. Efficient algorithms for mining outliers from large data sets. pages 427–438, 2000.

[136] R. H. Riedi, M. S. Crouse, V. J. Ribeiro, and R. G. Baraniuk. A multifractal wavelet model with application to network traffic. *IEEE Transactions on Information Theory*, 45(4):992–1018, 1999.

[137] J. B. T. M. Roerdink and M. A. Westenberg. Wavelet-based volume visualization. *Nieuw Archief voor Wiskunde*, 17(2):149–158, 1999.

[138] A. Ron. Frames and stable bases for shift invariant subspaces of l. *Canad. J. Math.*, (47):1051–1094, 1995.

[139] D. Roverso. Multivariate temporal classification by windowed wavelet decomposition and recurrent neural networks. In *In International Topical Meeting on Nuclear Plant Instrumentation, Controls, and Human-Machine Interface Technologies (NPIC&HMIT 2000)*, Washington, DC, November 2000.

[140] S. Santini and A. Gupta. A data model for querying wavelet features in image databases. In *Multimedia Information Systems*, pages 21–30, 2001.

[141] S. Santini and A. Gupta. Wavelet data model for image databases. In *IEEE Intl. Conf. on Multimedia and Expo*, Tokyo,Japan, August 2001.

[142] P. Scheunders, S. Livens, G. V. de Wouwer, P. Vautrot, and D. V. Dyck. Wavelet-based texture analysis. *International Journal on Computer Science and Information Management, ijcsim*, 1(2):22–34, 1998.

[143] C. Shahabi, S. Chung, M. Safar, and G. Hajj. 2d TSA-tree: A wavelet-based approach to improve the efficiency of multi-level spatial data mining. In *Statistical and Scientific Database Management*, pages 59–68, 2001.

[144] C. Shahabi, X. Tian, and W. Zhao. TSA-tree: A wavelet-based approach to improve the efficiency of multi-level surprise and trend queries on time-series data. In *Statistical and Scientific Database Management*, pages 55–68, 2000.

[145] G. Sheikholeslami, S. Chatterjee, and A. Zhang. WaveCluster: A multi-resolution clustering approach for very large spatial databases. In *Proc. 24th Int. Conf. Very Large Data Bases, VLDB*, pages 428–439, 1998.

[146] G. Sheikholeslami, A. Zhang, and L. Bian. A multi-resolution content-based retrieval approach for geographic images. *GeoInformatica*, 3(2):109–139, 1999.

[147] P. Smyth. Probabilistic model-based clustering of multivariate and sequential data. In *Proceedings of Artificial Intelligence and Statistics*, 1999.

[148] E. J. Stollnitz, T. D. DeRose, and D. H. Salesin. Wavelets for computer graphics: A primer, part 1. *IEEE Computer Graphics and Applications*, 15(3):76–84, 1995.

[149] E. J. Stonllnitz, T. D. DeRose, and D. H. Salesin. *Wavelets for computer graphics, theory and applications*. Morgan Kaufman Publishers, San Francisco, CA, USA, 1996.

[150] G. Strang. Wavelets and dilation equations: A brief introduction. *SIAM Review*, 31(4):614–627, 1989.

[151] G. Strang. Wavelet transforms versus fourier transforms. *Bull. Amer. Math. Soc.*, (new series 28):288–305, 1990.

[152] V. Strela. Denoising via block wiener filtering in wavelet domain. In *3rd European Congress of Mathematics, Barcelona*. Birkhauser Verlag., July 2000.

[153] Z. R. Struzik and A. Siebes. The haar wavelet transform in the time series similarity paradigm. In *Proceedings of PKDD'99*, pages 12–22, 1999.

[154] Z. R. Struzik and A. Siebes. Measuring time series' similarity through large singular features revealed with wavelet transformation. In *DEXA Workshop 1999*, pages 162–166, 1999.

[155] S. R. Subramanya and A. Youssef. Wavelet-based indexing of audio data in audio/multimedia databases. In *IW-MMDBMS*, pages 46–53, 1998.

[156] G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5), July 2002.

[157] G. Tzanetakis, G. Essl, and P. Cook. Automatic musical genre classification of audio signals. pages 205–210, Bloomington, IN, USA, October 2001.

[158] P. Vaidyanathan. Multirate digital filters, filter banks, polyphase networks, and applications: a tutorial. In *Proc. IEEE*, number 78, pages 56–93, 1990.

[159] M. Vannucci and F. Corradi. Covariance structure of wavelet coefficients: theory and models in a bayesian perspective. *J. R. Statist. Soc. B*, (61):971–986, 1999.

[160] R. Venkatesan, S. Koon, M. Jakubowski, and P. Moulin. Robust image hashing. In *Current proceedings*, 2000.

[161] B. Vidakovic. Nonlinear wavelet shrinkage with Bayes rules and Bayes factors. *Journal of the American Statistical Association*, 93(441):173–179, 1998.

[162] B. Vidakovic. Wavelet-based nonparametric bayes methods. Technical report, ISDS, Duke University, 1998.

[163] B. Vidakovic. *Statistical Modeling by Wavelets*. John Wiley & Sons, New york, 1999.

[164] J. S. Vitter and M. Wang. Approximate computation of multidimensional aggregates of sparse data using wavelets. pages 193–204, 1999.

[165] J. S. Walker. *A Primer on Wavelets for Their Scientific Applications (Studies in Advanced Mathematics)*. CRC Press, 1999.

[166] J. Z. Wang, G. Wiederhold, and O. Firschein. System for screening objectionable images using daubechies' wavelets and color histograms. In *Interactive Distributed Multimedia Systems and Telecommunication Services*, pages 20–30, 1997.

[167] J. Z. Wang, G. Wiederhold, O. Firschein, and S. X. Wei. Content-based image indexing and searching using daubechies' wavelets. *International Journal on Digital Libraries*, 1(4):311–328, 1997.

[168] Y. Wang. Jump and sharp cusp detection by wavelets. *Biometrika*, 82(2):385–397, 1995.

[169] P. Wojtaszczyj. *A Mathematical Introduction to Wavelets*. Cambridge University Press, Cambridge, 1997.

[170] P. C. Wong and R. D. Bergeron. Authenticity analysis of wavelet approximations in visualization. In *IEEE Visualization*, pages 184–191, 1995.

[171] B. J. Woodford and N. K. Kasabov. A wavelet-based neural network classifier for temporal data. In *Presented at the 5th Australasia-Japan Joint Workshop*, University of Otago, Dunedin, New Zealand, November 2001.

[172] Y.-L. Wu, D. Agrawal, and A. E. Abbadi. A comparison of DFT and DWT based similarity search in time-series databases. In *CIKM*, pages 488–495, 2000.

[173] T. Yamakawa, E. Uchino, and T. Samatsu. Wavelet neural network employing over-complete number of compactly supported non-orthogonal wavelets and their applciations. In *Proc. of 1994 IEEE International Conference Joint on Neural Networks*, volume 3, pages 1391–1396, 1994.

[174] R. Young. *Wavelet Theory and its Application*. Kluwer Academic Publishers, Bonston, 1993.

[175] A. Zhai, P. Huang, and T. J. yu Pan. A study on web-log using wavelet. In *Research and Development in Information Retrieval*, 2001.

[176] T. Zhang, R. Ramakrishnan, and M. Livny. BIRCH: an efficient data clustering method for very large databases. In *Proceedings of ACM SIGMOD*, pages 103–114, 1996.

[177] S. Zhu, T. Li, and M. Ogihara. CoFD: An algorithm for non-distance based clustering in high dimensional spaces. In *Proceedings of 4th International Conference DaWak 2002*, number 2454 in LNCS, pages 52–62, Aix-en-Provence, France, 2002. Springer.

Appendix A: Formal Definition

A function $\psi(x) \in L^2(R)$ is called a *wavelet* if it satisfies the following properties:

- $\int \psi(x)dx = 1$;
- There is a finite interval $[a, b]$, such that $\psi(x) = 0$ for all $x \notin [a, b]$;
- There exists a function $\phi(x)$ such that $\langle \phi, \psi \rangle = 0$ (i.e., ϕ is orthogonal to ψ) and satisfies

$$\phi(x) = \sum_{i=0}^{n} h_i \phi(2x - i),$$

for some $h_i, i = 0, n$;

- There exist a finite sequence of real numbers $g_0, \ldots, a_n$ such that

$$\psi(x) = \sum_{i=0}^{n} g_i \phi(2x - i);$$

- The dyadic dilation and translation of ϕ,

$$\psi_{j,k} = 2^{-j/2}\psi(2^j x - k), \quad \text{where} \quad j, k \in Z$$

is an orthonormal basis.

Appendix B: More on Fast DWT Algorithm

For a given function $f \in L^2(R)$ one can find a N such that $f_N \in V_N$ approximates f up to predefined precision (in terms of L^3 closeness). If $g_i \in W_i, f_i \in V_i$, then $f_N = f_{N-1} + g_{N-1} = \sum_{i=1}^{M} g_{N-i} + f_{N-M}$. Informally, we can think of the wavelets in W_i as a means of representing the parts of a function in V_{i+1} that can not be represented in V_i. So the decomposition process is as follows: given a f_N in V_N, we first decompose f_n into two parts where one part is in V_{N-1} and the other part is in W_{N-1}. At next step, we continue to decompose the part in V_{N-1} obtained from previous step into two parts where one in V_{N-2} and the other in W_{N-2}. This procedure is then repeated. This is exactly the wavelet decomposition.

Recall that we have $\phi(x) = \sum_{k=-\infty}^{\infty} a_k \phi(2x - k)$ and $\psi(x) = \sum_{k=-\infty}^{\infty} (-1)^k \bar{a}_{1-k} \phi(2x - k)$ where $\phi(x)$, the scaling function, is related to the space V_0 and $\psi(x)$, the mother wavelet function, is related to W_0. Define $b_k = (-1)^k a_{1-k}$ and usually the sequences $\{a_k\}, \{b_k\}$ are called Quadrature Mirror filters(QMF) in the terminology of signal processing. a_k is a low-band or low-pass filter and b_k is a hi-band or hi-pass filter. For a sequence $f = \{f_n\}$ that represents the discrete signal to be decomposed and the operators H and G are defined by the following coordinativewise relations:

$$(Hf)_k = \sum_n a(n - 2k)f(n), (Gf)_k = \sum_n b(n - 2k)f(n)$$

(This can be represented as convolution: $Hf = f(k) * a(n - k), Gf = f(k) * b(n-k)$). They represent filtering a signal through digital filters $a(k), b(k)$ that corresponds to the mathematical operation of convolution with the impulse response of the filters. The factor $2k$ represents downsampling. The operators H and G correspond to one-step in the wavelet decomposition. Thus the DWT transformation can be summarized as a single line:

$$\begin{aligned} f &\rightarrow (Gf, GHf, GH^2 f, \ldots, GH^{j-1} f, H^j f) \\ &= (d^{(j-1)}, d^{(j-2)}, \ldots, d^{(0)}, c^{(0)}), \end{aligned}$$

where we can call $d^{(j-1)}, d^{(j-2)}, \ldots, d^{(1)}, d^{(0)}$ coefficients details and $c^{(0)}$ coefficient approximation. The details and approximation are defined by iterative way: $c^{(j-1)} = Hc^{(j)}, d^{(j-1)} = Gd^{(j)}$.

Acknowledgment

The authors would like to thank the anonymous reviewers for their invaluabale comments. This work was supported in part by NSF Grants EIA-0080124, DUE-9980943, and EIA-0205061 and by NIG Grants 5-P41-RR09283, RO1-AG18231, and P30-AG18254.

About the Authors

Tao Li received his BS degree in Computer Science from Fuzhou University, China and MS degree in Computer Science from Chinese Academy of Science. He also got a MS degree in mathematics from Oklahoma State University. He is currently a doctoral candidate in the computer science department at University of Rochester. His primary research interests are: data mining, machine learning and music information retrieval.

Qi Li received his BS degree from Department of Mathematics, Zhongshan University, China in 1993, a Master degree from Department of Computer Science, University of Rochester in 2002. He is currently a PHD student in Department of Computer and Information Science, University of Delaware. His current interests are visual data mining and object recognition.

Shenghuo Zhu obtained a bachelor degree in Computer Science at Zhejiang University in 1994, and a master degree in Computer Science at Tsinghua University in 1997. He has been pursuing his Ph.D degree in the Computer Science Department at University of Rochester since 1997. His primary research interests are machine learning, data mining and information retrieval.

Mitsunori Ogihara received a PhD in Information Sciences at Tokyo Institute of Technology in 1993. He is currently Professor and Chair of the Department of Computer Science at the University of Rochester. His primary research interests are data mining, computational complexity, and molecular computation.