Privacy Preserving Technique for Euclidean Distance Based Mining Algorithms Using a Wavelet Related Transform

Mohammad Ali Kadampur and Somayajulu D V L N

National Institute of Technology (NITW),
Warangal, A.P. India 506004
{kadampur,soma}@nitw.ac.in
http://www.nitw.ac.in

Abstract. Privacy preserving data mining is an art of knowledge discovery without revealing the sensitive data of the data set. In this paper a data transformation technique using wavelets is presented for privacy preserving data mining. Wavelets use well known energy compaction approach during data transformation and only the high energy coefficients are published to the public domain instead of the actual data proper. It is found that the transformed data preserves the Eucleadian distances and the method can be used in privacy preserving clustering. Wavelets offer the inherent improved time complexity.

Keywords: Privacy, Data Mining, Wavelet Transforms.

1 Introduction

Data perturbation is one of the well known privacy preserving techniques[2]. It refers to a data transformation process typically performed by the data owners before publishing their data. Owners achieve two goals by transforming the data. First,the data gets disguised and sensitive information is not available to the public domain. Second,such transformations best preserve all those domain specific data properties that are critical for building meaningful data mining models[1,11]. These modified data sets or data models maintain task specific data utility of the published data. The tasks may vary from simple statistical analysis to the hidden knowledge discovery. The models built from data perturbation techniques are useful for applications where data owners want to participate in cooperative mining but at the same time want to prevent the leakage of privacy sensitive information in their published datasets[3,7,8,10]. In this paper we try to build such data models using a wavelet transform technique.

1.1 A Background of Wavelets

The wavelet transform of a wavelet $\psi(x)$ is mathematically defined as[6,12]:

$$W(a,b) = \int_x f(x) \frac{1}{\sqrt{a}} \psi\left(\frac{x-b}{a}\right) \tag{1}$$

C. Fyfe et al. (Eds.): IDEAL 2010, LNCS 6283, pp. 202–209, 2010.

which means for every (a, b) we will have a wavelet transform coefficient, representing how much the scaled wavelet is similar to the function at location $x = \frac{a}{b}$. Wavelet transform basically quantifies the local matching of the wavelet with the signal. If the wavelet matches the shape of the signal well at a specific scale and location then a large transform value is obtained otherwise if the wavelet and the signal do not correlate well, a low value of transform is obtained. Essentially application of any wavelet on a signal involves obtaining correlated coefficients as the wavelet slides along the signal[1,11]. The value of the coefficients depends on the wavelet chosen.

1.2 Wavelet Decomposition

The discrete wavelet transform(DWT) is an implementation of the wavelet transform using a discrete set of the wavelet scales and translations obeying some defined rules[7,6]. In other words, this transform decomposes the signal into mutually orthogonal set of wavelets, which is the main difference from the continuous wavelet transform (CWT), or its implementation for the discrete time series sometimes called discrete time continuous wavelet transform (DT-CWT). The key advantage of it is temporal resolution: it captures both frequency and location information (location in time).

In this paper we apply a variant of Haar[6] wavelet transformation. The Haar transform decomposes the input data into approximation coefficients and detailed coefficients. These coefficients in fact correspond to the low frequency and high frequency decompositions of the original samples respectively. The wavelet literature is rich with many interesting ways of such decomposition and reader is encouraged to refer[6,12].

2 Our Approach

In the proposed approach we treat entire data set as a centralized data set and design an algorithm that transforms the data set into a synthetic data set. The approach is based on the following observations

1. For most of the real data sets the energy of each transformed record is represented by very few coefficients.
2. High energy coefficient in one transformed record may have low energy coefficient in some others, but on an average energy tends to concentrate in a small set of transform coefficients. Therefore it is just sufficient if we identify and publish only such high energy coefficients.

The objective of the algorithm is to generate a set of coefficients for each record and select a set of high energy coefficients across a large number of transformed records. The method is illustrated with an example in the following section.