

Improving Medical/Biological Data Classification Performance by Wavelet Preprocessing

Qi Li
Department of CIS
University of Delaware
Newark, DE 19716
qili@cis.udel.edu

Tao Li , Shenghuo Zhu
Department of CS
University of Rochester
Rochester, NY 14620
{taoli,zsh}@cs.rochester.edu

Chandra Kambhampettu
Department of CIS
University of Delaware
Newark, DE 19716
chandra@cis.udel.edu

Abstract

Many real-world datasets contain noise and noise could degrade the performances of learning algorithms. Motivated from the success of wavelet denoising techniques in image data, we explore a general solution to alleviate the effect of noisy data by wavelet preprocessing for medical/biological data classification. Our experiments are divided into two categories: one is of different classification algorithms on a specific database (*Ecoli* [6]) and the other is of a specific classification algorithm (decision tree) on different databases. The experiment results show that the wavelet denoising of noisy data is able to improve the accuracies of those classification methods, if the localities of the attributes are strong enough.

1. INTRODUCTION

Noise is a random error or variance of a measured variable [3]. Many real-world datasets contain noise. There are many possible reasons for noisy data, such as measurement errors during the data acquisition, human and computer errors occurring at data entry, technology limitations and natural phenomena. 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. Most of these methods, however, are not specially designed in order 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 have been successfully applied in image research area. The main idea of wavelet denoising is to transform the data into the wavelet domain, 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. Though wavelet techniques have been widely used for image data, little work has been reported on using wavelet techniques to denoise other kinds of data, say, medical/biological data which are mainly obtained by experiments or measurements and hence have a good chance of containing noise. This is because image data usually have strong (spatial) locality¹, but locality of medical/biological data is usually hidden. Although medical/biological data lack the spatial locality, in our recent investigation, we found that most medical/biological data contain a certain kind of locality which makes the use of wavelet techniques for denoising plausible. Take the *Ecoli* database [6] for example: the *Ecoli* database is used for predicting the cellular localization sites of proteins and it contains 336 instances with 8 attributes for each instance. The 6th attribute represents the score of discriminant analysis of the amino acid content of outer membrane and periplasmic proteins. It originally contains 8 classes where one class has 5 instances, two other classes have 2 instances each, 3 classes are subcases for a big class and 2 other classes are subcases of another big class. So we then simply the 8 classes into 4 classes: cytoplasm, periplasm, inner membrane and outer membrane. We organize the data according to these four classes, i.e., the data in the same class are placed together and plot the distribution of their 6th attribute as shown in Figure 1. We observe good locality of the data from Figure 1 and hence it is plausible to use wavelet techniques to remove the noise.

2. WAVELET DENOISING

¹By locality, we refer to continuity in the sense that the variance of the data is relatively small in its neighborhood.

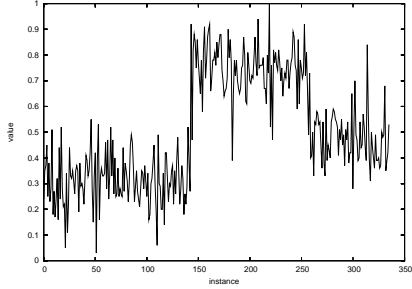


Figure 1: Locality of the 6th attribute in the *Ecoli* dataset

Suppose observation data $y = (y_1, \dots, y_n)$ is a noisy realization of the signal $x = (x_1, \dots, x_n)$, $y_i = x_i + \epsilon_i$, $i = 1, \dots, n$, 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), $MSE(\hat{x}) = \frac{1}{n} \sum_{i=1}^n (\hat{x}_i - x_i)^2$.

[2] has 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. There are three commonly used shrinkage functions: the hard, soft and the non-negative garrote shrinkage functions:

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

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

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

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 threshold* λ^* is defined as threshold λ which minimizes expression

$$\inf_{\lambda} \sup_{\theta} \left\{ \frac{R_\lambda(\theta)}{n^{-1} + \min(\theta^2, 1)} \right\}, \quad (2.1)$$

where $R_\lambda(\theta) = E(\delta_\lambda(x) - \theta)^2$, $x \sim N(\theta, 1)$. Interested readers can refer to [7] for other methods.

3. EXPERIMENTAL RESULTS

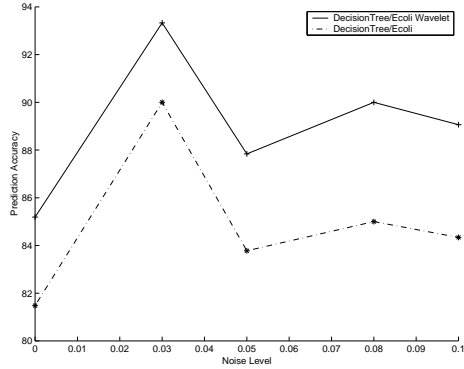
We use the minimax threshold to denoise since it has been reported to be very efficient. We also choose wavelet Db4 [1] in our experiments. The complete experimental results description can be found in our tech report [4].

3.1 Wavelet denoising for different classifiers

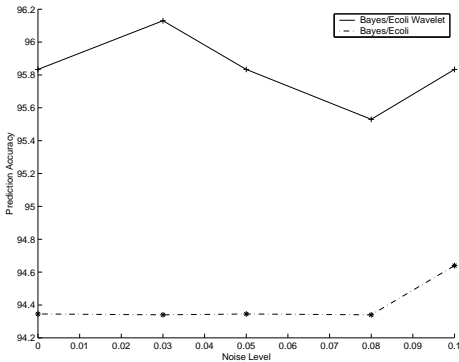
In this section, we investigate the effects of wavelet denoising for different classification techniques on *Ecoli* database which is available from UCI machine learning repository [6]. We compare the performance of different classifiers on *Ecoli* database between two cases: data with wavelet preprocessing and data without wavelet preprocessing. To further demonstrate the effects of wavelet techniques, we also derive several additional database by injecting noise into the *Ecoli* database and perform the comparison across a range of noise levels.

The four classifiers we used are: decision tree, naive Bayes, PART rule learner, and oneR. The *Ecoli* database is used to predict the cellular localization sites of proteins and it contains 336 instances with 8 attributes for each instance. We organized the data according to the 6th attribute which represents the score of discriminant analysis of the amino acid content of outer membrane and periplasmic proteins and it may contain noise due to measurement errors. To denoise the database, we then preprocess data of 6th column with waveShrink technique before applying the classifiers on the database. We also derive 4 additional noisy databases from the original *Ecoli* database by injecting different levels of noise. Denote the standard deviation of the 6th attribute of the original *Ecoli* database as σ_6 . The noise we add into the *Ecoli* database satisfy Gaussian distributions with zero mean and standard deviations: percentage $\times \sigma_6$, $i = 1, 2, 3, 4$, where percentage $_i = 0.03, 0.05, 0.08$ and 0.1 .

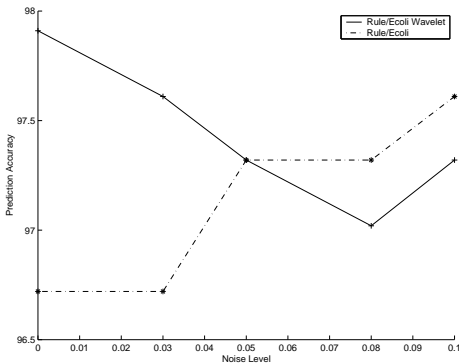
Figure 2 shows the comparisons between classification accuracies on data with wavelet preprocessing and data without preprocessing by decision tree, naive Bayes classifier, PART rule learner and oneR respectively. The accuracies are obtained by three-fold cross-validation. The improvements in decision tree, naive Bayes and oneR are obvious. There are about 1.6% improvement in decision tree, 1.4% improvement in naive Bayes and 3% improvement in oneR on average. And at different noise levels, the performance of decision tree, naive Bayes classifier and oneR on data with wavelet preprocessing is always better than those without preprocessing. The performance of PART rule learner on data with preprocessing before noise level 0.05 is better than those without preprocessing and after noise level 0.05, the former is beaten by the latter. This half-part success of wavelet preprocessing obviates our experience and intuition which tell us that the average performance



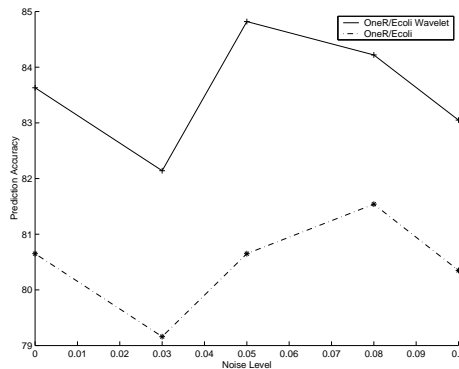
(a) decision tree



(b) naive Bayes



(c) PART rule learner



(d) oneR

Figure 2: The performance with four classifiers.

is usually degraded with the increasing noise levels. So we tend to see the performance comparison with PART rule learner on the *Ecoli* database a neutral result (neither positive nor negative).

In short, wavelet preprocessing is a promising method to improve the classification performance for different classifiers on the *Ecoli* database. But notice that the good locality (Figure 1) of the *Ecoli* database may not be owned by the other medical databases. So we need to study the effect of wavelet preprocessing on different kind of databases which is the task of Section 3.2.

3.2 Wavelet denoising on different databases

In this section, we investigate the effects of wavelet denoising on different databases with decision tree techniques. We compare the performance of decision tree classifiers on data with wavelet preprocessing and those without preprocessing. The databases we used are heart disease databases. They are also available in UCI machine learning depository [6]. We choose decision tree classifier since it is one of the most widely used techniques in practice.

Heart disease databases consist of real, experimental data from four international medical organizations, Cleveland Clinic Foundation, Hungarian Institute of Cardiology, the University Hospitals in Zurich and Basel in Switzerland, and V.A. Medical Center in Long Beach, California (VAMC). These databases have been widely used by researchers to develop prediction models for coronary diseases. There are a large amount of missing data of VAMC. 689 entries out of (200×13) are missing. Since missing values may seriously bias the threshold estimation, we will only consider the heart disease databases of *Cleveland*, *Hungarian* and *Switzerland*.

The arrangement of data is indexed by the monotonically increasing age of sampling people. The eighth column of the heart disease database is data on maximum heart rate. The maximum heart rate is affected by the activity and/or levels of fitness. The different activity may result a difference of 3 – 5 beats in the number. For example, studies show that maximum heart rate on a treadmill is consistently 5 – 6 beats higher than on a bicycle ergometer and 2 – 3 beats higher on a rowing ergometer [5]. Also improper procedures introduce errors in the measurement. We also derive 4 additional noisy databases from the original database by injecting different levels of noise. The performances are obtained by randomly splitting the dataset into two: 80% for training and 20% for testing. Figure 3 shows the performances of decision tree classifier on the three heart disease databases with or without preprocessing on the eighth column. The results on *Cleveland* and *Hungarian* are positive. The classification accuracies with wavelet preprocessing are

always higher than those without preprocessing on these two heart disease databases. The result on *Switzerland* is neutral. The classification accuracies are unchanged with preprocessing.

Similar results were observed on the fourth column of the heart databases. In addition, we did experiments on SPECTF and Pima Indians diabetes disease [6]. Readers could refer our tech report[4] for a complete description of all the experimental results.

4. CONCLUSIONS AND FUTURE WORK

In this paper, we study a general solution to reduce the noise sensitivity of by wavelet denoising. Two sets of experiments (on different classifiers and different databases) show that the preprocessing of noisy data is able to improve classification accuracies if the localities of attributes are strong enough.

Our experiments also show that lack of locality is the biggest hurdle for applying wavelet tool to the classification. Without locality or with weak locality, the wavelet domain (wavelet coefficients) is unable to characterize the noise accurately. Although most of medical/biological data have neither temporal locality nor spatial locality, it is still possible to arrange an artificial locality. For example, we may sort the data to bring continuity before wavelet denoising. We are currently exploring the techniques to discover/enhance localities of medical/biological data.

5. REFERENCES

- [1] I. Daubechies. Ten lectures on wavelets. SIAM, Philadelphia, 1992.
- [2] David L. Donoho and Iain M. Johnstone. Minimax estimation via wavelet shrinkage. *Annals of Statistics*, 26(3):879–921, 1998.
- [3] Jiawei Han and Micheline Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers, 2000.
- [4] Qi Li, Tao Li, and Shenghuo Zhu. Improving medical/biological data classification performance by wavelet preprocessing. Technical Report 788, July 2002.
- [5] Londeree and Moeschberger. Effect of age and other factors on hr max. *Research Quarterly for Exercise and Sport*, 53(4):297–304, 1982.
- [6] UCI. Machine learning databases. In <ftp://ftp.ics.uci.edu/pub/machine-learning-databases/>.
- [7] Hong ye Gao. Wavelet shrinkage denoising using the non-negative garrote. *Journal of Computational and Graphical Statistics*, 7(4):469–488, 1998.

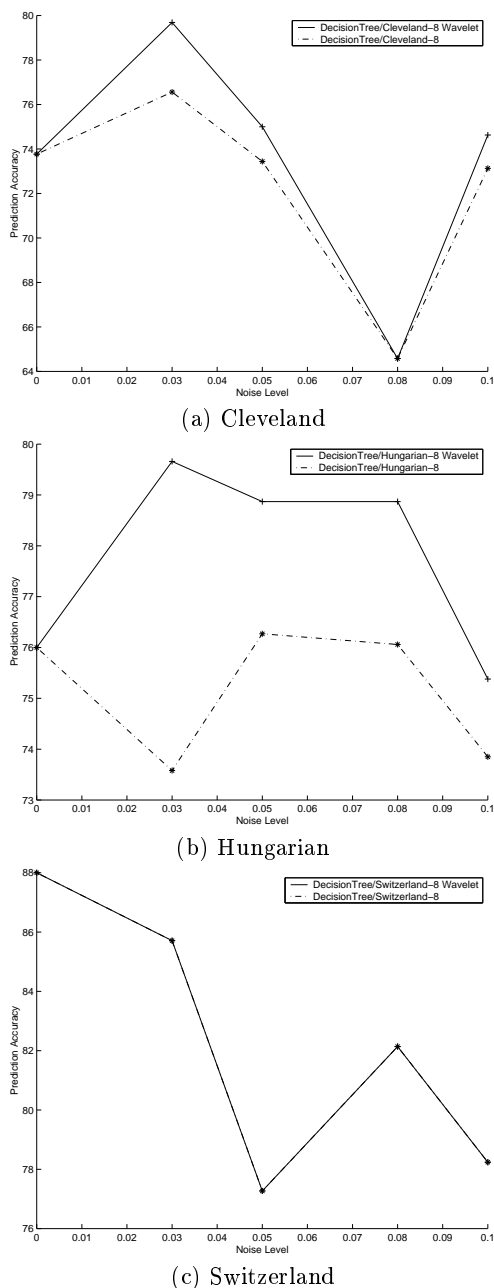


Figure 3: The performances on three heart disease databases.