A robust system for distributed data mining and preserving privacy

El Sayed Mahmoud and David Calvert
School of Computer Science, University of Guelph

Introduction
Interest in knowledge-based collaborative applications has emerged due to the availability of large volumes of data that can be analyzed through the Internet. Many organizations in several domains are motivated to combine their records to improve the reliability and completeness of the extracted knowledge. However, they cannot disclose their records for privacy reasons.

This work proposes a new robust multiple classifier system (MCS) called Gdadapt that recognizes patterns distributed across multiple sites while avoiding the transfer of individual records between those sites to preserve privacy. A classifier is built based on the local data of each site. All classifiers are transferred to a central site within the local data. These are potential classifiers for building the Gdadapt. A synthetic dataset is generated using these classifiers to be used as a validation dataset for Gdadapt.

The Purpose
This work aims to develop a robust MCS (Gdadapt) for preserving privacy. The focus of this work is to investigate the effect of using the synthetic dataset generated based on the potential classifiers on the performance of Gdadapt.

Contributions
• Developing a robust MCS (Gdadapt) for preserving privacy.
• Determining how to generate synthetic data that are used to preserve privacy and improve the performance of Gdadapt.

The data
The data used represent problems with multiple overlaps between class regions. A data generator is designed to produce datasets with predefined characteristics such as Bayes error. An example of a dataset with Bayes error 0.1 is shown in Figure 1.

Methods
Gdadapt consists of 3 layers as shown in Figure 2:
• Layer 1 uses the classifiers constructed based on the local data of sites to generate a synthetic dataset for validation.
• Layer 2 combines an analytic model for fitness, a multi-objective genetic algorithm to construct MCSs with optimized fitness. The analytic model decomposes the fitness into three components: accuracy, positive-diversity, and negative diversity. The accuracy and positive-diversity are used as two objectives for NSGAII.
• Layer 3 uses BKS to determine and assign a score for each local region. A local region is a subset of the feature space defined by the boundaries of the MCS members. The score of a region is the probability of a pattern located in a region being correctly identified by the region’s most frequent class.

Two methods for generating the synthetic dataset are investigated. The first method is to use a random classifier to label each generated pattern; the second is to use major voting. The performance of Gdadapt when using a real validation dataset is compared to the performance when using the two synthetic validation datasets.

Results
The performance of Gdadapt when using three different validation datasets 30 times is shown in Figure 3. The three validation datasets are: (1) real data, (2) synthetic data using random classifiers, and (3) synthetic data using major voting of all classifiers. Figure 4 shows the average of three performance differences between the performance when using the real and the two synthetic datasets are not significant since the significance level is greater than 0.05 as shown in Figure 5.

Conclusions
A new system for pattern recognition across multiple sites while preserving privacy was designed and tested. The design is based on combining the following: (1) an analytic model of fitness to provide a performance measure; (2) a multi-objective genetic algorithm to identify a diverse set of MCSs with optimized fitness; (3) a data generator to generate a synthetic validation data set; and (4) BKS to determine and score local regions of the MCSs identified in 3.

Two methods for generating the validation dataset were tested. The first is based on selecting a random classifier for each unlabeled synthetic pattern. The second method is based on applying major voting to the classifiers constructed at sites to label each pattern.

Using major voting of classifiers constructed at sites to label random patterns for generating a validation dataset for constructing the Gdadapt does not make a significant reduction in the performance of Gdadapt. This is a promising result for developing distributed data mining algorithms that preserve privacy.

Bibliography
1.E. Mahmoud and D. Calvert, “A novel analytic model for fitness of a majority vote multi-classifier”, Pattern Recognition, 2012 (Submitted on March 2012)