Our main research interests include data mining algorithms and their applications. For example, if you have a dataset with some missing and incorrect values, we can clean that up for you by imputing the missing values, identifying (and making corrections of) the corrupt values, and carrying out various other pre-processing tasks. The datasets cleaned up by our algorithms are more useful and accurate for various statistical and data mining analyses. Moreover, the algorithms automatically learn the properties of a dataset, identify (and make necessary corrections of) incorrect values, and impute/estimate missing values without requiring any user input and domain knowledge. Of course with user input and domain knowledge we can carry out additional cleaning, but that is not a requirement.

We also analyse data and extract patterns from them through our classification algorithms that build decision trees and decision forests from datasets. The extracted patterns will help you to understand your datasets better. While understanding your data through conventional statistical analyses such as correlation calculation you need to assume the presence or absence of a relationship between two attributes/features such as Productive Employees and Salary, our pattern extraction (knowledge discovery) algorithms do not require any assumptions. Instead they automatically find such relationships and logic rules. You simply need to give them your questions such as “Why some employees are productive and some are not?” and the algorithms will come up with all possible answers and their statistical significances. With the discovered knowledge our algorithms can then also predict the future; whether or not a potential new employee will be productive. That sounds interesting!

We are also interested in clustering that can find useful groups of records (such as customers and patients) having similar properties. Our in house clustering algorithms do not require any user input whatsoever including the number of clusters. They also allow you to put different weights (if you wanted) varying from 0 to 1 on the attributes to indicate the significance levels of the attributes for your clustering purpose.

Privacy Preserving Data Mining is another research focus of us. If you want to release your dataset for public use, but are concerned about the privacy of the data subjects our privacy preserving techniques will allow you to add noise to the datasets for preserving the privacy while maintaining the quality of the data. Another recent research focus is the possible threats from data mining on the privacy of online social network site users and their technical solutions.

Each dataset comes with its own challenges and requirements. For example, some datasets are very unstable in nature having high dimension (a few thousand of attributes) and low size (only few records). Some datasets are very imbalanced in the sense that they have huge number of records (say 99.99% of the total records) of one class/group such as Non-Cancer and only few records (say 0.01%) of the other class such as Cancer. The datasets can also be time series, sequential and tabular having various types of attributes including categorical, numerical, binary, nominal and ordinal. Since we develop in house and custom-made algorithms, we can cater for your domain specific requirements and challenges. We are also interested to help you to design your survey questions in order to build a useful dataset.

We have applied our algorithms in irrigation water demand prediction, software defect prediction, property price prediction and employee management. For all these real world problems our algorithms were found very useful. Will they also be useful in analysing yours?