Tutorials
ICDM 2008

T1 – Integration of Classification and Pattern Mining: A Discriminative and Frequent Pattern-based Approach

Tutorialists
Hong Cheng, Chinese University of Hong Kong
Jiawei Han, Univ. of Illinois at Urbana-Champaign
Xifeng Yan, IBM T. J. Watson Research Center
Philip S. Yu, University of Illinois at Chicago

Abstract
Classification is a core method widely studied in machine learning, statistics, and data mining. Many classification methods assume that the input data is in a feature vector representation. However, in many tasks, the predefined feature space is not discriminative enough to distinguish different classes. More seriously, in many other applications, the input data, such as those in the form of transactions, sequences, graphs, and semi-structured data, may have no clear, predefined feature vectors. In both scenarios, a primary challenge is how to construct a discriminative and compact feature set. There have been considerable research efforts for classifying transaction, sequence, and graph data. Recent researches have reported some promising results on such data with the newly developed frequent-pattern based classification. However, there has been no systematic tutorial on such a methodology. This tutorial presents a comprehensive, organized, and state-of-the-art survey on methodologies and algorithms on mining discriminative patterns for classification, with an emphasis on several aspects: accuracy, efficiency, interpretability, and application. This contributes to a systematic introduction to this new methodology that integrates classification with pattern mining, and therefore, may benefit research progress in both frontiers. This tutorial is prepared for data mining, machine learning and statistics researchers who are interested in classifying both structured and unstructured data.

T2 – Privacy-Preserving Location Services

Tutorialist
Mohamed F. Mokbel, University of Minnesota, Twin Cities

Abstract
The explosive growth of location-detection devices (e.g., GPS-like devices and handheld devices) along with wireless communications and mobile databases results in realizing location-based applications that deliver specific information to their users based on their current locations. Examples of such applications include location-based store finder, location-based traffic reports, and location-based advertisements. Although location-based services promise safety and convenience, they threaten the privacy and security of users as such services explicitly require users to share private location information with the service and possibly with others. If a user wants to keep her location information private, she has to turn off her location-aware device and temporarily unsubscribe from the service. Unfortunately, recent studies show that such privacy concerns – ranging from worries over employers snooping on their workers’ whereabouts to fears of tracking by potential stalkers – are a serious obstacle to wider
T3 – Sample Selection Bias – Covariate Shift: Problems, Solutions, and Applications

Tutorialists
Wei Fan, IBM T. J. Watson Research
Masashi Sugiyama, Tokyo Institute of Technology

Abstract
Sample selection bias/covariate shift is a common problem encountered when using data mining algorithms for many real-world applications. Traditionally, it is assumed that training and test data are sampled from the same probability distribution, the so-called “stationary or non-biased distribution assumption.” However, this assumption is often violated in reality. Typical examples include marketing solicitation, fraud detection, drug testing, loan approval, school enrollment, medical diagnosis etc. For these applications the only labeled data available for training is a biased representation, in various ways, of the future data on which the inductive model will predict. Intuitively, some examples sampled frequently into the training data may actually be infrequent in the testing data, and vice versa. When this happens, an inductive model constructed from biased training set may not be as accurate on unbiased testing data if there had not been any selection bias in the training data. For example, there has been speculations that the most recent US subprime mortgage problem is due to sample selection bias problem where the default customers do not follow the same risk model as traditional mortgage customers. In this tutorial, we will employ various examples to describe the problem, describe various solution, and end the tutorial with a systematic approach to address a real-world problem.

T4 – Mining Ubiquitous Data Streams: From Theory to Applications

Tutorialists
Joao Gama, University of Porto
Shonali Krishnaswamy and Mohamed Medhat Gaber, Monash University

Abstract
Many sources produce data continuously. Examples include customer click streams, wireless sensors (e.g., bio sensors, environmental sensors etc.), telephone records, web logs, multimedia data, and retail chain transactions. These datasets of continuous and distributed nature are called data streams. Processing distributed data streams is of paramount importance in the research community, as new algorithms are needed to process this streaming data in real time. The goal of this tutorial is to present the state-of-the-art in mining data streams and discuss open research problems, issues, and challenges in this area. We will present techniques for change detection, clustering, classification, frequent patterns, and time series analysis from distributed data streams. Real-world applications and case studies in the areas of Intelligent Transportation Systems, Patient Monitoring and Habitat Monitoring will be presented in detail stimulating the real need for this growing research field. Finally, the tutorial will be concluded with open issues and future directions. The tutorial also provide a substantial list of data stream mining resources.