Abstract

Statistical Machine Learning techniques offer approaches to designing programs where there is considerable uncertainty in the data due to noise and finite sample sets. The talk will begin with a tutorial overview of some formalism in machine learning, including the Bayesian approach, graphical models, principal components, and Gaussian mixture models. Examples of their application in computational forensics where the goal is to match input evidence to known exemplars will be given. Specific examples from signature verification, fingerprint matching and shoe-print matching will be described.
Abstract

Binary data have been occupying a special place in the domain of data analysis. Analysis of binary data sets, however, generally leads to NP-complete/hard problems. Consequently, the focus here is on effective heuristics for reducing the problem size. Multilinear algebra, the algebra of higher-order tensors, offers a potent mathematical framework for analyzing the multifactor structure of high dimensional real world data. The multilinear modeling technique employs a tensor extension of the conventional matrix singular value decomposition (SVD), nonnegative matrix decomposition (NMF), semi discrete decomposition (SDD) etc.

There are several well known methods and algorithms for factorization of real data but many application areas including information retrieval, pattern recognition and data mining require processing of binary rather than real data. Unfortunately, the methods used for real matrix factorization fail in the latter case. In this lecture we introduce background for binary matrix factorization. In order to perform object recognition (no matter which one) it is necessary to learn representations of the underlying characteristic components. Such components correspond to object-parts, or features. These data sets may comprise discrete attributes, such as those from market basket analysis, information retrieval, and bioinformatics, as well as continuous attributes such as those in scientific simulations, astrophysical measurements, and sensor networks. The feature extraction if applied on binary datasets, addresses many research and application fields, such as association rule mining, market basket analysis, discovery of regulation patterns in DNA microarray experiments, etc. So called bars problem is used as the benchmark. Set of artificial signals generated as a Boolean sum of given number of bars is analyzed by these methods. Here we will concentrate on the case of black and white pictures of bars combinations represented as binary vectors, so the complex feature extraction methods are unnecessary. Generally we can employ the lattice theory within the field of computational intelligence.

Many applications in computer and system science like as clustering, classification, pattern analysis and matrix decomposition, involve analysis of large scale and often high dimensional lattice data. Therefore, suitable methods approximating the data in lower dimensions or with lower rank are needed. In the following, we focus on the factorization of high-dimensional binary or lattice data or high order binary or lattice tensors.
Abstract
The concept of Information Fusion was initiated by the need of using several data sources and technical terms like data fusion and multisensor fusion were born. This concept intends to give a global framework to pattern recognition systems using several sensors. Some important examples raised rapidly: Target Identification Systems, Parcels Recognition in remote sensing, Medical Diagnostic Systems, etc. The use of existing theories like Bayesian, Belief Functions, Fuzzy and Possibility and the real world applications have certainly stimulated new research activities and new engineering challenges.

For instance: available information is not limited to simple data but includes knowledge sources (not necessarily numerical); different information sources may be affected by different forms of imperfections (missing data, uncertainty, ambiguity, etc.). These questions show clearly that information fusion is not the simple concatenation of exiting theories and that we need a global and homogeneous framework where:

- the basic concept of information is clearly defined;
- the mathematical imperfections models are defined and linked to imperfection sources affecting the observed, or available information;
- different existing theories and their scope of use and application are clearly positioned.

This invited talk will tackle these important questions. A precise definition of the concept of an information element is given; different sources of information imperfections are presented, and, mathematical characterization of information elements is expressed so that the different information fusion technical approaches are positioned. Pattern recognition systems mainly using imaging sensors are used to clarify and give concert examples of the proposed global approach leading to consider Information Fusion as a single theory where existing technical approaches constitute the components of this theory.
Intelligent Wood Species Recognition

Marzuki Khalid  
University Teknologi Malaysia  
Malaysia

Abstract

Tropical rainforest has more than 3,000 different types of timber species. According to the Forest Research Institute of Malaysia, out of these about 200 species are being used by the timber industry. Among the major timber consumers are housing developers, wood fabricators and furniture manufacturers where the need for recognition of wood species is necessary. Automatic wood recognition has not yet been well established mainly due to lack of research in this area and the difficulty in obtaining the wood database. In this paper, an automatic wood recognition system based on image processing, feature extraction and artificial neural networks was designed. A proto-type portable wood recognition machine has been developed which is capable of classifying 30 different tropical Malaysian woods according to their species based on the macroscopic wood anatomy. Image processing is carried out using our newly developed in-house image processing library referred to as “Visual System Development Platform”. The textural wood features are extracted using a variety of feature extraction techniques such as the grey-level co-occurrence matrix, local binary pattern and linear discriminant analysis. To further improve the classification of the wood species not only the popular multi-layered neural network based on the back-propagation algorithm is used, other techniques such as support vector machine, wavelet packet decomposition and gabor filters are used. An intelligent decision module based on fuzzy inference is developed to increase the accuracy of the classification results. The device can provide wood identification within seconds, eliminating the need for laborious human recognition. The results obtained show a high rate of recognition accuracy proving that the techniques used are suitable to be implemented for commercial purposes.
Abstract

The problem of mining frequent patterns in relational databases has motivated many research efforts during the last decade. This motivation is due to the fact that frequent queries (i.e., queries whose answers contain a sufficient number of tuples) are useful for detecting hidden trends in a database. These trends are generally expressed as rules, called association rules (similarly to the popular association rules dealing with frequent itemsets), but also as constraints holding on the database, such as functional dependencies or inclusion dependencies. Although many efforts have been devoted towards this goal, in particular based on the so called Apriori algorithm, the general problem of mining frequent queries from a relational database is known to be untractable. This explains why all approaches dealing with this problem consider restricted frameworks in which such mining becomes feasible. These restrictions concern mainly the type of queries to be mined (all approaches focus on conjunctive queries) and the structure of database from which queries are mined (either a single table, or a database with multiple tables). During the talk, the main existing approaches will be briefly outlined and then, our current work will be presented in more details. Our approach deals with conjunctive queries in the following two settings: (i) a single relational table, and (ii) a relational database which operates on a star schema (star schemas are used for modeling data warehouses). The main feature of our approach is to exploit the constraints (i.e., the functional and inclusion dependencies) on the data in order to avoid redundant computations when mining frequent queries. More precisely, in either of the two cases mentioned above, queries are compared according to a pre-ordering with respect to which the support of queries is shown to be anti-monotonic (a property required to apply a strategy a la Apriori). Then, it turns out that, using the induced equivalence relation, only one query per equivalence class has to be considered, because equivalent queries are shown to have the same support. The impact of this basic property will be discussed and algorithms will be presented, along with their complexity analysis. Preliminary experiments will show that, based on our work, mining frequent conjunctive queries can be tractable.
Collaborative Security Mechanism in Detecting Intrusion Activity

Shahrin Sahib
Universiti Teknikal Malaysia Melaka
Malaysia

Abstract

The rapidly increasing array of Internet-scale threats is a pressing problem for every organization that utilizes the network. Organizations often have limited knowledge as well as capability to detect and respond to these threats. The sharing of information related to probes and attacks is a facet of an emerging trend toward “collaborative security”. Collaborative security mechanism provide network administrator with a valuable tool in this increasingly hostile environment.

Intrusion detection is an important part of network systems security protection and become a major device inside the collaborative network. IDS have the capabilities to analyze the network traffic and recognize incoming and on-going intrusion. Unfortunately the combination of fast attack and slow attack detection module for anomaly detection in real time may slow down the detection process. In real time network, early detection of fast attack detection can prevent any further attack and reduce the unauthorized access on the targeted machine. Current IDS researches more focus on detection technique using various techniques such as Neural Network and Data Mining without considering the suitable features. Feature such as IP address in the most important to identify the attacker and victim of the attack. Therefore, selecting the suitable set of feature selection and the correct threshold value, add an extra advantage for the IDS to detect the anomalies for fast attack detection.

Although fast attack detection increased the accuracy of detection but IDS still need further improvement in order to reduce the fast alarm. Therefore, correlating data among different logs to improve intrusion detection systems accuracy is needed. The existing alert correlation techniques had been reviewed and analyzed which are similarities based, Predefined Attack Scenarios, Pre-Requisites and Consequence of Individual Attack and Statistical Causal Analysis Technique. From the analysis, six capability criteria have been identified to improve the current alert correlation technique. They are capability to do alert reduction, alert clustering, identify multistep attack, reduce false alert, detect known attack and detect unknown attack. Therefore, the comparison of various techniques needs to be analyzed to find the most suitable log correlation technique for detecting the attackers. The accurate log generated by IDS using fast attack and alert correlation technique can further use as a source of evidence for forensic investigation in order to trace the origin of potential attack. Therefore, the research on IDS will help to increase the accuracy of the intrusion detection and become an input for forensic research in tracing the potential attackers.
Abstract
The problem of mining frequent patterns in relational databases has motivated many research efforts during the last decade. This motivation is due to the fact that frequent queries (i.e., queries whose answers contain a sufficient number of tuples) are useful for detecting hidden trends in a database. These trends are generally expressed as rules, called association rules (similarly to the popular association rules dealing with frequent itemsets), but also as constraints holding on the database, such as functional dependencies or inclusion dependencies. Although many efforts have been devoted towards this goal, in particular based on the so called Apriori algorithm, the general problem of mining frequent queries from a relational database is known to be untractable. This explains why all approaches dealing with this problem consider restricted frameworks in which such mining becomes feasible. These restrictions concern mainly the type of queries to be mined (all approaches focus on conjunctive queries) and the structure of database from which queries are mined (either a single table, or a database with multiple tables). During the talk, the main existing approaches will be briefly outlined and then, our current work will be presented in more details. Our approach deals with conjunctive queries in the following two settings: (i) a single relational table, and (ii) a relational database which operates on a star schema (star schemas are used for modeling data warehouses). The main feature of our approach is to exploit the constraints (i.e., the functional and inclusion dependencies) on the data in order to avoid redundant computations when mining frequent queries. More precisely, in either of the two cases mentioned above, queries are compared according to a pre-ordering with respect to which the support of queries is shown to be anti-monotonic (a property required to apply a strategy a la Apriori). Then, it turns out that, using the induced equivalence relation, only one query per equivalence class has to be considered, because equivalent queries are shown to have the same support. The impact of this basic property will be discussed and algorithms will be presented, along with their complexity analysis. Preliminary experiments will show that, based on our work, mining frequent conjunctive queries can be tractable.