Advanced Topics in Data Mining

Instructor:
Dr. Hamid Turab Mirza

Department of Computer Science
CUI, Lahore

INTRODUCTION

Outline

- What is Data Mining
- How does Data Mining differ from other approaches?

Introduction

The aim of data mining is

1) to make sense of

2) large amounts of

3) mostly unsupervised data, in some

4) domain

Introduction

1) to make sense

we envision that new knowledge should exhibit a series of essential attributes:

be understandable

valid

novel

useful

Introduction

1) to make sense

the most important requirement is that the discovered new knowledge needs to be

***understandable* to data owners**

who want to use it to some advantage

Introduction

1) to make sense

a model that can be described in easy-to-understand terms, like production rules such as:

IF abnormality (obstruction) in coronary arteries

THEN coronary artery disease

Introduction

1) to make sense

the second most important requirement is that the discovered new knowledge should be *valid*

If all the generated rules were already known the rules would be considered trivial and of no interest

(although the generation of the already-known rules validates the generated model)

Introduction

1) to make sense

the third requirement is that the discovered new knowledge must be ***novel***

If the knowledge about how to diagnose a patient had been discovered not in terms of but, say, a neural network then this knowledge may or may not be acceptable, since a neural network is a “black box” model.

A trained neural network might still be acceptable if it were proven to work well on hundreds of new cases.

Introduction

1) to make sense

the fourth requirement is that the discovered new knowledge must be *useful*

Usefulness must hold true regardless of the type of model used.

Introduction

2) large amounts

DM is about analyzing large amounts of data that cannot be dealt with by analyzing them manually.

Introduction

2) large amounts

AT&T handles over 300 million calls daily to serve about 100 million customers and stores the information in a multiterabyte database

Wal-Mart, in its stores handles about 21 million transactions a day, and stores the information in a database of about a dozen terabytes

NASA generates several gigabytes of data per hour through its Earth Observing System

Introduction

2) large amounts

Oil companies like Mobil Oil store hundreds of terabytes of data about different aspects of oil exploration

The Sloan Digital Sky Survey project will collect observational data of about 40 terabytes

Modern biology creates, in projects such as human genome/proteome, data measured in terabytes and petabytes

Homeland Security is collecting petabytes of data on its own and other countries’ citizens.

Introduction

3) mostly unsupervised data

It is much easier, and far less expensive, to collect unsupervised data than supervised data.

For supervised data we must have known inputs that correspond to known outputs, as determined by domain experts.

Introduction

3) mostly unsupervised data

What can we do if only unsupervised data are collected?

Use algorithms that are able to find “natural” groupings/clusters or relationships/associations in the data.

If clusters are found they can be possibly labeled by domain experts. If we are able to do it unsupervised data becomes supervised and the problem becomes much easier.

Introduction

3) mostly unsupervised data

What to do when the data are semisupervised (meaning that there are a few known training data pairs along with thousands of unsupervised data points)?

Can these few data points help in the process of making sense of the entire data set?

There exist techniques, called

***semi-supervised learning*,**

which take advantage of these few training data points, for instance partially supervised clustering

Introduction

3) mostly unsupervised data

A DM algorithm that works well on both small and large data is called

scalable

unfortunately, few are.

Introduction

4) domain

The success of DM projects depends heavily on access to domain knowledge.

Discovering new knowledge from data is a highly interactive (with domain experts)

and iterative (within knowledge discovery) process.

We cannot take a successful DM system, built for some domain, and apply it to another domain and expect good results.

Introduction

The ultimate goal is to provide students with the fundamentals of frequently used DM methods and to guide them in their DM projects:

from understanding the problem and the data,

through preprocessing the data,

to building models of the data and

validating these

to putting the newly discovered knowledge to use.

Introduction

www.kdnuggets.com

This web site is by far the best source of information about all aspects of DM.

How does Data Mining Differ from Other Approaches?

Data mining came into existence in response to technological advances in many disciplines:

Computer Engineering contributed significantly to the development of more powerful computers in terms of both speed and memory;

Computer Science and Mathematics continued to develop more and more efficient database architectures and search algorithms;

and the combination of these disciplines helped to develop the World Wide Web (WWW).

How does Data Mining Differ from Other Approaches?

Along with dramatic increase in the amount of stored data came demands for better, faster, cheaper ways to deal with those data.

In other words,

all the data in the world are of no value without mechanisms to efficiently and effectively extract

information/knowledge

from them.

How does Data Mining Differ from Other Approaches?

Early DM pioneers:

U. Fayyad

H. Mannila

G. Piatetsky-Shapiro

G. Djorgovski

W. Frawley

P. Smith

many others…

How does Data Mining Differ from Other Approaches?

Data mining is not just an “umbrella” term coined for the purpose of making sense of data.

In statistics, researchers frequently deal with the problem of finding the smallest data size that gives sufficiently confident estimates.

In DM, we deal with the opposite problem, namely, data size is large and we are interested in building a data model that is small (not too complex) but still describes the data well.

How does Data Mining Differ from Other Approaches?

Finding a good model of the data, which at the same time is easy to understand,

is at the heart of DM.

We need to keep in mind that none of the generated models will be complete (using all the relevant variables/attributes of the data),

and that almost always we will look for a compromise between model completeness and model complexity.

How does Data Mining Differ from Other Approaches?

Word of caution:

although many commercial as well as open-source DM tools exist they

do not by any means produce automatic results

in spite of the vendors hype about them.

The users should understand that the application of even a very good tool to one’s data will most often not result in the generation of valuable knowledge for the data owner after simply clicking “run”.

References

Cios, K.J., Pedrycz, W., and Swiniarski, R. 1998. *Data Mining Methods for Knowledge Discovery*. Kluwer

Han, J., and Kamber, M. 2006. *Data Mining: Concepts and Techniques*. Morgan Kaufmann

Hand, D., Mannila, H., and Smyth, P. 2001. *Principles of Data Mining*. MIT Press

Hastie, T., Tibshirani, R., and Friedman, J. 2001. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer

Kecman, V. 2001. *Learning and Soft Computing*. MIT Press

Witten, H., and Frank, E. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann