Tomáš Horváth

INTRODUCTION TO DATA MINING

Lecture 2

CRISP-DM

Institute of Computer Science, Faculty od Science

Pavol Jozef Šafárik University in Košice

Slovak Republic

Overview

The aim of this lecture is to introduce You the CRISP-DM methodology[1] in more details[2].

[1] http://www.crisp-dm.org

[2] Pete Chapman (NCR), Julian Clinton (SPSS), Randy Kerber (NCR), Thomas Khabaza (SPSS), Thomas Reinartz (DaimlerChrysler), Colin Shearer (SPSS) and Rüdiger Wirth (DaimlerChrysler): CRISP-DM 1.0 - Step-by-step data mining guide, 2000.

About the CRISP-DM

A methodology developed in the project[1] (number 24.959), partially funded by the European Commission under the ESPRIT Program.

Project partners

- NCR Systems Engineering Copenhagen[2], USA and Denmark.
  - Data warehouse
- SPSS Inc.[3], USA.
  - Data mining solutions.
- DaimlerChrysler AG[4], Germany.
  - car industry
- OHRA Verzekering en Bankk Groep B.V.[5], Netherlands
  - insurance industry

[1] http://www.crisp-dm.org
[2] http://www.ncr.com
[3] http://www.spss.com
[4] http://www.daimlerchrysler.com
[5] http://www.ohra.nl

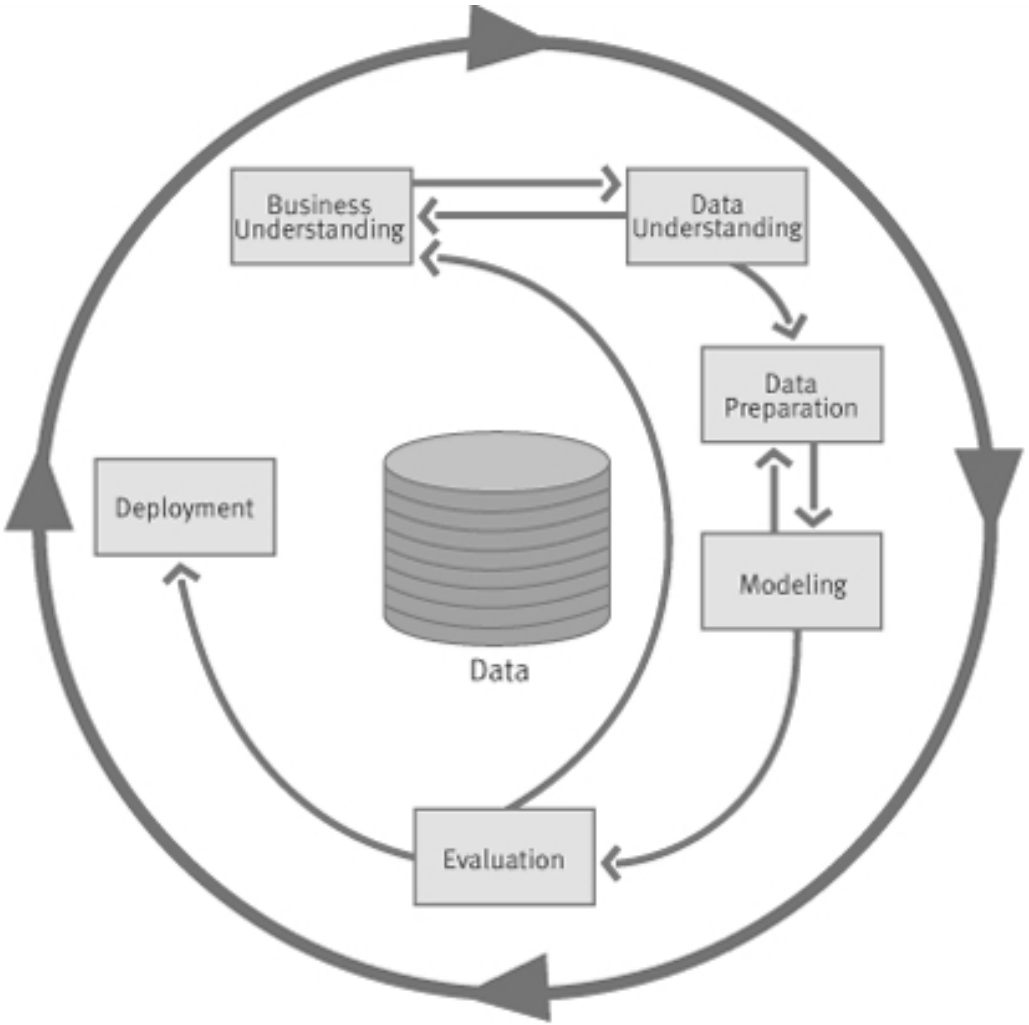
What is CRISP-DM?

CRoss **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining

Four levels of abstraction

- Phase
  - The data mining process is organized into several phases consisting of **tasks**.
- Generic task
  - The general level for tasks which should be **complete** (covering the whole data mining process as all possible applications) and **stable** (valid for yet unforeseen techniques).
- Specialized task
  - Description of tasks in certain specific situations, how they will be provided, etc.
- Process Instance
  - The record of actions. decisions and results of the actual data mining engagement.

The Process model



[1]

[1] Image source: http://www.crisp-dm.org

I. Business Understanding

The aim is to understand the needs of a client, the requirements and business objectives, convert the objectives to data mining goals, uncover important factors influencing these outcomes and prepare a preliminary plan for achieving the goals.

Generic tasks of this phase are

1. Determine business objectives
   - understanding the client's needs from the business perspective[1]
2. Assess situation
   - investigation of facts about the factors influencing the project
3. Determine data mining goals
   - determining the project objectives in technical terms[2]
4. Produce project plan
   - preparation of a detailed plan to reach the project objectives

[1] e.g. "Increase catalog sales to our customers."

[2] "Predict how many things customers will buy given information collected about them."

I.1. Determine business objectives

Goal: **Background**

- *Collate information about the client's business situation, identify human and material resources which could be used.*

Deliverables:

- Organization
  - divisions, departments, key persons and their responsibilities
  - a steering committee of the project
- Problem
  - the problem in general
  - the current status and prerequisities (motivation, usage of DM)
  - users' needs, the project results (i.e. written report for top management, running system for users)
- Current solution
  - advantages, disadvantages and the acceptance feedback of the current solution (if any)

I.1. Determine business objectives

Goal: **Business objectives**

- *Describe the primary objective and – in addition – the secondary objectives from a business perspective.*

Deliverables:

- the problem to be solved
- as precise specificaton of all business questions and other business requirements as possible
- expected benefits in business term

Beware of setting unrealistic goals!

I.1. Determine business objectives

Goal: **Business success criteria**

- *Describe the success and usefulness criteria for the project outcome from a business perspective in a quite specific and measurable terms.*

Deliverables:

- business success criteria specification (e.g. improve customers' response rate by 15%)
- persons assessing the criteria

Each criteria should relate to some specified business objective(s).

I.2. Assess situation

Goal: **Inventory of resources**

- *List the available data, software, hardware and human resources which can be used in the project.*

Deliverables:

- Hardware
  - hardware, its availability and maintenance schedule for the project as well as its adequacy for the DM tools to be used (if known)
- Data, Knowledge and Tools
  - data and knowledge sources as their type (on-line, experts, hand-written)
  - available tools and the relevant background knowledge
- Personnel
  - system admin, database admin, tech support and other staff
  - market analysts, domain and DM experts and their availability

Before starting this task, consider previous experiences with this – or similar – problem(s).

I.2. Assess situation

Goal: **Requirements, assumptions and constraints**

- *List the requirements of the project including the schedule, quality of results, security and legal issues as well as assumptions on data usage and the constraints of the project.*

Deliverables:

- Requirements
  - scheduling, accuracy, deployment, maintainability and repeatability
  - security, legal restrictions, privacy, reporting schedule
- Assumptions
  - data quality, external factors and cost estimates, reporting type
- Constraints
  - legal issues, budget, timescales, resources
  - rights to data sources, (technical) accessibility of data and relevant knowledge

The list of assumptions should also include assumptions determined at the beginning of the project.

I.2. Assess situation

Goal: **Risks and contingencies**

- *List the risks and the corresponding contingencies for recovering from the occurence of risks and mitigating their impact to the project.*

Deliverables:

- Risks
  - business, organizational, financial and technical risks as well as risk depending on data quality
- Contingencies
  - triggers of risks and the corresponding conditions as well as contingency plans

I.2. Assess situation

Goal: **Terminology**

- *Write a glossary of business as well as data mining terminology relevant to the project.*

Deliverables:

- prior availability of glossaries
- domain experts' terminology
- business terminology

I.2. Assess situation

Goal: **Costs and benefits**

- *Prepare a cost-benefit analysis of the project.*

Deliverables:

- costs for data collection
- costs of the solution (development and implementation)
- benefits from the solution
- operating costs

One should try to prepare as specific comparison (costs-benefits) as possible.
Identify also hidden costs (e.g. repeated data extraction, changes in schedule, training).

I.3. Determine data mining goals

Goal: **Data mining goals**

- *Describe the intended technical outputs of the project enabling to achieve the business objectives.*

Deliverables:

- business questions in data mining terminology and data mining problem type(s)

Goal: **Data mining success criteria**

- *Define success criteria in subjective, technical terms.*

Deliverables:

- criteria for model assessment
- benchmarks for evaluation

I.4. Produce project plan

Goal: **Project plan**

- *Prepare the detailed plan of the project including gantt chart, dependencies, milestones and risks.*

Deliverables:

- initial process plan and its feasibility to all participants
- identified goals, selected techniques
- effort and resources needed
- critical steps
- decision and review points
- major iterations

I.4. Produce project plan

Goal: **Initial assessment of tools**

- *Perform an initial assessment of data mining tools and techniques for different stages of the process.*

Deliverables:

- list of selection criteria for tools and techniques for each phase of the process
- potential tools and techniques, their reviews and evaluation of their appropriateness

II. Data Understanding

The aim is to collect initial data, get familiar with data and identify the quality of data as well as detect subsets interesting to form some hypotheses.

Generic tasks of this phase are

1. Collect initial data
   - acquisition of data listed in resources and understanding them as well as initial data preparatin steps
2. Describe data
   - examination of the surface properties of acquired data
3. Explore data
   - querying, visualization and reporting data directly addressing the data minng goals
4. Verify data quality
   - examination of the quality of data

II.1. Collect initial data

Goal: **Initial data collection report**

- *List data used within the project, define importance of attributes and identify problems of merging data.*

Deliverables:

- Data requirements planning
  - information needed and the availability of these information
- Selection criteria
  - identified selection criteria (necessary and irrelevant attributes, amount of data suitable for a choosen technique, . . .)
  - tables of interest and data within these tables
  - length of the history the data will be used in
- Insertion of data
  - encoding of free text entries, methods for acquisition of missing attributes, data extraction mechanisms

Keep in mind possible inconsistencies in merged data.
Some important information/knowledge sources about the data may be non-electronic.

II.2. Describe data

Goal: **Data description report**

- *Describe acquired data including the format, the quantity, the identifiers of the fields and other discovered features.*

Deliverables:

- Volumetric analysis of data
  - methods of data capture, data source access, statistical analyzes, tables and their relations, data volume and redundancies
- Attribute types and values
  - accessibility and availability of attributes, their types, value ranges, correlations, meanings of attributes in business terms, basic statistics and their analyzes, relevancy to specific goals (together with the domain expert), balancing the data (if necessary)
- Keys
  - key relations and overlap of key values
- Review assumptions
  - updated list of assumptions if necessary

II.3. Explore data

Goal: **Data exploration report**

- *Describe initial hypotheses and their impact, report data characteristics and detect interesting data subsets for further examination.*

Deliverables:

- Data exploration
  - analyze properties in more depth and detect interesting subsets in data
- Suppositions for further analyzis
  - evaluation of findings in the data description report
  - formed hypothesis and identified actions and their transformation to data mining goals
  - clarified data mining goals
  - basic analyzis to verify hypotheses

II.4. Verify data quality

Goal: **Data quality report**

- *Describe the quality of data and list possible solutions for emerged problems.*

Deliverables:

- Special values and their meaning
  - keys and coverage
  - coincidence of meanings of attributes and contained values
  - identified missing and blank values as well as their meanings
  - attributes with similar meanings but different values
  - deviations and if these are noise or not, plausability of values
  - consistencies of delimiters and number of fields in flat files
  - consistencies and redundancies
  - type of noise and how to deal with it

Make reviews on conflicting attributes.
Use visualization to better look on the data.

III. Data Preparation

The aim is to construct the final dataset from raw data which will be the input for the modeling tool.

Generic tasks of this phase are

1. Select data
   - decision on the data used for analysis according to their relevance to the specified objectives
2. Clean data
   - improve the quality of data as the selected analysis techniques require
3. Construct data
   - perform constructive data preparation operations
4. Integrate data
   - integrate data from multiple tables
5. Format data
   - mainly syntactic modifications of data to be suitable for the modeling tools

III.1. Select data

Goal: **Rationale for inclusion/exclusion**

- *List the data to be included in as well as excluded from the process and provide reasons for these decisions.*

Deliverables:

- appropriate additional data from different sources
- significance and correlation tests
- reconsidered data selection (task II.1.) in light of experiences on data quality, exploration and modeling
- selected different data subsets (data which meets certain conditions)
- available sampling techniques (if the tool can't handle the full dataset)
- documentation of rationale for inclusion/exclusion

One can weight the attributes accordingly to their importance.

III.2. Clean data

Goal: **Data cleaning report**

- *Reconsider the decisions and actions from the Verify Data Quality task (II.4.) to detect which issues are still out-standing and what affect these can have to the outcome of the project.*

Deliverables:

- reconsidered noise-handling procedures
- corrected, removed or ignored noise
- treatment with special values and their meaning
- reconsidered data selection (task II.1.) in light of experiences on data cleaning

If the noise is ignored for some attributes (because is irrelevant) it should be documented.

III.3. Construct data

Goal: **Derived attributes**

- *Derive new attributes constructed from one or more other, existing attributes because (i) from the background knowledge we know that some additional facts might be important, (ii) the data mining technique used handles only certain types of data or (iii) we have a hunch that certain facts were not covered.*

Deliverables:

- normalized and transformed attribute values (if needed)
- added "relevance" of attributes (as new attributes)
- imputed/completed missing values according to the decided type of construction
- added the derived attributes to data

Before deriving an attribute one should check how it eases the model.

III.3. Construct data

Goal: **Generated records**

- *Add completely new records to the data which represent new knowledge (e.g. the prototypes of data segments).*

Deliverables:

- available techniques needed
- added new records

III.4. Integrate data

Goal: **Merged data**

- *Join tables having different information about the same objects and, also, generate new records and aggregated values if reasonable.*

Deliverables:

- checked integration facilities for their ability to integrate
- integrated database
- reconsidered data selection criteria (task II.1.) in light of experiences of data integration

Remember for sources in non-electronic format.

III.5. Format data

Goal: **Reformatted data**

- *Modify (primarily) syntactically to fulfill the format required by the data mining tool used.*

Deliverables:

- rearranged attributes
- reordered records
- reformatted within-values

IV. Modeling

In this phase, modeling techniques are selected and their parameters are tuned to optimal values.

Generic tasks of this phase are

1. Select modeling technique
   - selection of the actual modeling technique
2. Generate test design
   - generation of a procedure to validate the model and test it's quality
3. Build model
   - run the modeling technique to build models
4. Assess model
   - interpretation, evaluation, comparison and ranking of models according to the evaluation criteria from a data mining perspective

IV.1. Select modeling technique

Goal: **Modeling technique**

- *Record an actually used modeling technique.*

Deliverables:

- approppriate technique chosen according to the tool selected

Goal: **Modeling assumptions**

- *Record specific assumptions for an actually used modeling technique.*

Deliverables:

- assumptions about the data for an actually used modeling technique and their comparison with the Data description report (task II.2.)

IV.2. Generate test design

Goal: **Test design**

- *Describe plans for training, testing and evaluating the model.*

Deliverables:

- existing test designs for the data mining goals
- necessary steps (iterations, folds, . . .)
- prepared test data

IV.3. Build model

Goal: **Parameter settings**

- *Set initial parameters and document reasons for the chosen values.*

Goal: **Models**

- *Run the selected technique and post-process the results.*

IV.3. Build model

Goal: **Model description**

- *Describe the resulting model and assessment of its properties.*

Deliverables:

- characteristics of the model and its parameter settings
- detailed description of the model with technical informations
- the interpretation of the model
- conclusions regarding possible patterns in data

IV.4. Assess model

Goal: **Model assessment**

- *Summarize the results of this task.*

Deliverables:

- test results of models acquired, their comparison and interpretation
- best models selected and their interpretation in business terms
- comments from domain experts on reliability, plausability, usefulness and novelty as well as impacts of these models
- analysis of potentials of deployment of these models
- insights why a certain modeling technique leads to good/bad results

Goal: **Revised parameter settings**

- *Adjust parameters to lead to better results.*

V. Evaluation

In this phase, the model is thoroughly evaluated to be certain that it achieves the business objectives, the whole process is reviewed and next steps are determined.

Generic tasks of this phase are

1. Evaluate results
   - evaluation of the achievments of business objectives
2. Review of process
   - summarization of the whole process and detecting important factors which could be overlooked
3. Determine next steps
   - decision of the next steps to be made

V.1. Evaluate results

Goal: **Assessment of data mining results with respect to business success criteria**

- *Summarize results in terms of business success criteria and make a final statement if the project achieved the business objectives.*

Deliverables:

- understanding and interpretation of results in terms of the application domain
- the impact, novelty and usefulness of the data mining result
- evaluation and comparison of results with respect to business success criteria
- new business objectives to be addressed later in the project
- conclusions for next data mining projects

Goal: **Approved models**

- *Get approved models which meet the stated criteria.*

V.2. Review process

Goal: **Review of process**

- *Summarize the process review and determine missed activities and steps which should be repeated.*

Deliverables:

- overview of the data mining process
- possible improvements
- failures and misleading steps
- possible alternative actions and unexpected paths

V.3. Determine next steps

Goal: **List of possible actions**

- *List possible further actions.*

Deliverables:

- analysis of the deployment
- estimation of improvement of the current process
- available remaining and additional resources
- recommendation of alternative continuations
- refined process plan

Goal: **Decisions**

- *Describe how to proceed.*

Deliverables:

- rank of possible actions
- reasons for the selected action

VI. Deployment

In this phase, the knowledge gained during the process is organized, eventually, presented for the customers.

Generic tasks of this phase are

1. Plan deployment
   - creation of the strategy for deployment of the project results into the business
2. Plan monitoring and maintenance
   - preparation of the maintenance strategy
3. Produce final report
   - final documentation of the project
4. Review project
   - experience documentation

VI.1. Plan deployment

Goal: **Deployment plan**

- *Summarize a detailed deployment strategy on the base of evaluation reports.*

Deliverables:

- deployable results and alternative plans for deployment
- summarized information and knowledge
- deployment propagation plan within the organization
- possible problems
- measures of benefits from the use of the results
- deployment plan

VI.2. Plan monitoring and maintenance

Goal: **Monitoring and maintenance plan**

- *Summarize a detailed monitoring and maintenance strategy.*

Deliverables:

- dynamic aspects which could change
- measurable criteria for stopping to use the model
- possible changes in business objectives with use of the model
- monitoring and maintenance plan

It is important to fully document the initial problem the model was intended to solve.

VI.3. Produce final report

Goal: **Final report**

- *Prepare a detailed final report including all the threads, describing the whole process, the costs, deviations from the original plans and recommendations for the future work.*

Goal: **Final presentation**

- *If necessary, prepare a final presentation for the management on the base of the final report.*

It is important to know the audience the report or the presentation is made to.

VI.4. Review project

Goal: **Experience documentation**

- *Summarize experiences gained during the project.*

Deliverables:

- experiences of significant members of the project
- end users' opinions and feedback
- analyse of the process (what have we learned, sensitive actions, critical points, . . .)

Try to abstract from details to make experiences useful for the future.


That's all Folks!
Thanks for your attention

Questions?

Tomas.Horvath@upjs.sk

http://www.ics.upjs.sk/~horvath