Application of data mining techniques
Course project description

Adam Zagdański

February 28, 2024

Contents

1 Project goal

- The main goal of the project is to use data mining methods (introduced during the course) to perform a complete analysis of selected data, related to a specific practical problem.
- An important aspect of the project should be performance evaluation and comparison of all methods/algorithms used in the analysis. Equally important are comprehensive conclusions regarding practical usefulness of employed data mining techniques in the context of problem considered.
- The project is divided into two main components

part I

1.) Descriptive analysis + data visualization (exploratory data analysis),

2.) Classification along with detailed accuracy assessment,

part II

3.) Cluster analysis with quality assessment,

4.) Application of the selected dimension reduction method in connection with classification and cluster analysis.

- Data sets that can be used in the project are listed in Chapter 2. In justified cases and after consulting the instructor, it would be possible to select data from outside this list.
- Details regarding the expected content and the structure of the project are given in Chapter 3.
- Additional information on methods and algorithms can be found in Chapter 4.
- In Chapter 5 you can find information about important dates (deadlines).
- In Chapter 6 additional remarks concerning the project are provided.

2 Datasets

We select **one** dataset from the following list to be analysed in the project:

A) **Medical diagnostics:** Breast Cancer Wisconsin (Diagnostic) Data Set
http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)
(Remark: please use data stored in wdbc.data file)

B) **Medical diagnostics:** Hepatitis Data Set
http://archive.ics.uci.edu/ml/datasets/Hepatitis

C) **Credit risk:** Statlog (German Credit Data) Data Set
http://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data)

D) **Socioeconomic study:** Adult Data Set
http://archive.ics.uci.edu/ml/datasets/Adult

E) **Automobile Data Set**
http://archive.ics.uci.edu/ml/datasets/Automobile

F) **Proteomics/SELDI-TOF technology:** Arcene Data Set
http://archive.ics.uci.edu/ml/datasets/Arcene.

G) **Medical diagnostics/DNA microarray data: microarray_datasets.zip**
One dataset can be chosen from: Leukemia data, Colon data, Prostate data, Lymphoma data, SRBCT data, Brain data.

H) **Spam filtering**
http://archive.ics.uci.edu/ml/datasets/Spambase
Note that this dataset is also available in the R (see `spam{ElemStatLearn}.`).

I) **Telco Customer Churn (the loss of customers problem):**
https://www.kaggle.com/datasets/blastchar/telco-customer-churn.

J) **Direct mail marketing response problem: dataset Clothing_Store.**
A description of the data can be found e.g. in the book T.Larose, Data Mining Methods and Models, Wiley & Sons.

K) **Bank marketing**
http://archive.ics.uci.edu/ml/datasets/Bank+Marketing.

Remark:
Some data contain many instances (records) and may cause problems related to the computational burden or effectiveness of methods. Therefore (in justified cases) a subset can be randomly selected for actual analysis. If you select such subset, please remember to describe it in detail in your report!

3 Expected content and structure of the project

List of **mandatory elements** that should be included in the project

- **Problem description and research questions formulation**
  Short information on the specificity of the problem considered. What questions do we want to answer analysing the data? What potential benefits may result from the analysis? For example, the benefit could be: a better diagnostic method, better efficiency in detecting bad/good customers applying for a loan, separating groups of customers who can be targeted with a specific offer, identifying relevant features/variables, etc.

- **Data characteristics**
  Data size, number of cases and features, types of features, information about missing values, information on unusual values (e.g. non-standard coding of missing values, etc.).

- **Methods and algorithms used in the project**
  What methods/algorithms have been used? Which tasks these methods/algorithms were used for? (e.g. preliminary analysis and visualization (exploratory data analysis), classification, prediction, cluster analysis).

- **Results**
  Results presented in the form of corresponding tables, graphs and diagrams. Note that only the most important results should be included in the report, whereas additional (supplementary) results can be added as attachments (e.g. PDF file containing additional figures).

- **Conclusions**
  Precise conclusions: what can be concluded from the analyses carried out? How these conclusions could be put into practice? (e.g. development of a new/better strategy in the company, new/better diagnostic method, etc.).

- **Further research suggestions**
  Short information on further possible directions of research (what could/should be further studied and what additional methods/algorithms could be used?)

Remark
It is recommended for the project structure to be in line with the basic **IMRAD** standard (Introduction, Methods, Results And Discussion), see e.g. http://en.wikipedia.org/wiki/IMRAD.

4 Methods and algorithms

- Appropriate methods/algorithms related to given data mining tasks should be chosen and used in the project.

  · **Descriptive analysis and data visualization**

    ◇ Objective

      ∗ Basic characteristics of variables/features (e.g. range, properties of distribution, etc.).

      ∗ Analysis of dependencies (correlation) of features.

      ∗ Initial assessment of discriminative ability of consecutive features (i.e. ability to separate objects from different classes).

      ∗ Identification of missing values and outliers.

    ◇ Methods/tools

      ∗ Summary statistics (measures of location and dispersion, computed for all data and within groups/classes)

      ∗ Basic charts (histograms, scatterplots, boxplots, etc.)

  · **Classification**

    ◇ Objective: construction of the classification rule (decision rule).

    ◇ Selected methods/algorithms, including: linear and quadratic discriminant analysis (LDA, QDA), k-nearest neighbors (k-NN), other approaches.

    ◇ Assessment of classification accuracy for various combinations of features (feature subsets) and different classification methods

      ∗ basic version – comparison of classification error using the training and test set split,

      ∗ advanced version – use of cross-validation or appropriate bootstrap-based procedure.

  · **Cluster analysis (clustering)**

    ◇ Objective: group objects according to their similarity.

    ◇ Selected methods/algorithms, including: k-means, PAM, AGNES and other.

    ◇ Quality assessment of cluster analysis results.

      ∗ basic version – comparison of average silhouette index values for different number of clusters K,

      ∗ advanced version – other internal indices assessing separation, compactness, etc. (including your own ideas!)

  · **Dimension reduction**

    ◇ Objective: feature extraction, visualization of multidimensional data.

    ◇ The use of selected methods (e.g. PCA or MDS) in connection with the classification and clustering.

- **Additional remarks:**
  - Please do not use methods/algorithms in an automated manner! In particular, one should check for which types of features a given method can be used and, possibly, precede its use with the selection of appropriate features or the use of appropriate pre-processing.
  - Note that for cluster analysis we cannot use a grouping variable (i.e. class labels) in order to allow the algorithm to discover the real data structure!
  - You are also encouraged to include additional methods in the project, e.g. non-standard classification and clustering algorithms, algorithms for discovering association rules, more advanced feature selection methods, etc.
  - It would be good if subsequent stages of the analysis were not treated as completely independent parts. For example, it is worth investigating whether the results of cluster analysis confirm the division into classes given by a specific grouping variable, etc.

5 Important dates and deadlines

- **Selection of a project topic / dataset (see chapter 2):** information should be given to the instructor in class or sent by e-mail by **15.03.2024** at the latest.
- **Part I (see chapter 1):** the first report (pdf + R/Python source files) should be submitted by **5.05.2024**.
- **Part II (see chapter 1):** the second report (pdf + R/Python source files) should be submitted by **21.06.2024**.
- **Remark:** Delay in submitting the project will result in getting a lower grade!

6 Remaining comments

- The project can be done individually or in pairs. However, when the project is prepared by two people the analysis should be sufficiently detailed and cannot be limited only to basic (the simplest) methods or algorithms.
- Do not forget to attach to the project (but do not paste!) the script(s) written in the R or Python language.
- After loading the data into workspace, please check whether all variable types have been correctly recognized (e.g. in R environment, variables of the type numeric and factor).
- The scope of the project is defined quite generally. However it is important to precisely address the problem and use appropriate data mining techniques in the analysis.
- You are encouraged to be inventive and creative! In case of any doubts you can always consult your ideas with the instructor.
- A few examples of non-standard analyses
  - Including the cost matrix in the assessment of the classification accuracy
  - Issues related to feature transformation
    - Discretization of continuous features (e.g. instead of the continuous variable `INCOME` one may use corresponding categorical variable `INCOME.CLASS` which takes value (category) from the set {'small', 'medium', 'large'}).
    - The construction of derivative features basing on the raw data (feature interaction). For example, instead of (or in addition to) features `AGE` and `GENDER` we use the derivative feature `AGE`×`GENDER`.
- The originality of the analysis and non-standard ideas will be rewarded!