Assignment 3: Clustering

Individual Home Assignment: Clustering
Published Online: 7 Nov 2016 (CHANGELOG12 DECEMBER 2016)
SUBMISSION DEADLINE: SUNDAY 15 JAN 2017, 23:59

Assignments' Deadlines
18 Dec 2016: Ass1 and Ass2
15 Jan 2017: Ass 3
24 Feb 2017: Final submission date for all assignments.

Learning objectives

In this assignment you are going to:

- Use the k-Means and Hierarchical clustering as implemented in Weka to perform unsupervised classification and exploration of the text categories included in the Swedish national corpus.

NB: In this assignment, you are required to select the machine learning methods and the options to be used in the tasks by yourselves, without step-by-step instructions. By now, you are familiar with the algorithms we studied in the course and you should be able to orientate yourselves in weka.

Data

The SUC datasets
Download the datasets on to your computer:
< http://stp.lingfil.uu.se/~santinim/ml/2016/Datasets/SUC_datasets/ >

The Stockholm-Umeå Corpus (or SUC) is the Swedish national corpus. The SUC is a collection of Swedish texts from the 1990's, consisting of one million words.
The original weka SUC dataset was recently created by Johan Falkenjack using readability features[1]. This original dataset was then divided into several subsets for carry out a number of experiments of text classification[2].
The SUC contains 500 samples of texts with a length of about 2,000 words each. Technically speaking, the SUC is divided into 1040 bibliographically distinct text chunks, each assigned to a category and a subsubgenre. The SUC contains nine top categories and 48 subcategories.

Dataset names are self-explanatory. Each dataset contains the same number of readability feature (ie 118 features), but a different number of classes and texts. See the breakdown of the datasets in Table 1. Capital letters indicate SUC text categories: see Table2 in the Appendix for the full list of domains, genres and subgenres.

[1] See Falkenjack et al. (2013). Features indicating readability in Swedish text. In Proceedings of the 19th Nordic Conference of Computational Linguistics (NoDaLiDa-2013), Oslo, Norway
<http://stp.lingfil.uu.se/~santinim/ml/2016/Assignments/_SwedishReadabilityFeatures2013_ecp1385008.pdf >.

[2] See Falkenjack et al. (2016). An Exploratory Study on Genre Classification using Readability Features. The Sixth Swedish Language Technology Conference (SLTC) Umeå University, 17-18 November, 2016
< http://stp.lingfil.uu.se/~santinim/ml/2016/Assignments/_SLTC_2016_paper_19.pdf >.

SUCdataset_SubGenres_48_1040texts_118readabilityCues.arff	48 subgenres
SUCdataset_TopGenres_9_1040rows_118_ReadabilityCues.arff	9 genres (A, B, C, E, F, G, H, J, K)
SUCdataset_TopGenresWithoutMisc_8_895texts_118readabilityCues.arff	8 genres (without H)
SUCdataset_SelectedGenres_6_709texts_118readabilityCues.arff	6 selected genres (A, B, C, G, J, K)
SUCdataset_SelectedDomainsWithMisc_3_331texts_118readabilityCues.arff	2 domains + Miscellanous (E, F, H)
SUCdataset_SelectedDomains_2_186texts_118readabilityCues.arff	2 domains (E, F)
Table 1. Breakdown of the SUC datasets.	

Goal of the Assignment

The goal of this assignment is to explore to what extent k-Means and Hierarchical Clustering in combination with readability features make sense of SUC's text categories. Since clustering does not rely on labelled examples, it needs robust features capable of revealing sensible patterns in data.
The underlying assumption is that domain and genre are two different notions that are not represented by the same type of features.

The following theoretical distinctions is provided to distinguish the notions of genre and domain:

- **Domain** is a subject field. Domain refers to the shared general topic of a group of texts. For instance, ”Fashion”, ”Leisure”, ”Business”, ”Sport”, ”Medicine” or ”Education” are examples of broad domains. In text classification, domains are normally represented by topical features, such as content words.
- **Genre** is a more abstract concept. It characterizes text varieties on the basis of conventionalized textual patterns. For instance, an academic paper obeys to textual conventions that differ from the textual conventions of a tweet ; or a letter complies to conventions that are different from the conventions of an interview . Academic papers , tweets , letters , interviews are examples of genres. Genre conventions usually affect the organization of the documents (its rhetorical structure and composition), the length of the text, the syntax and the morphology (e.g. passive forms v.s. active forms), vocabulary richness, etc. In text classification, genres are often represented by features such as POS tags, character n-grams, or POS n-grams.

How do readability features work on the whole SUC (9 text categories), on SUC subcategories (48 classes), on the six genres, on the 2 domains, etc.? Run k-Means and Hierarchical Clustrering on all the datasets listed in Table 1 to explore the efficiency of readability features using unsupervised machine learning algorithms.

G tasks

Theoretical question:

Q1: Describe and comment the main differences between k-Means and hierarchical clustering. Avantages and disadvantages of both.

Part 1

Start weka and choose the Explorer interface. Work with the SUC datasets.
Cluster the SUC using kMeans and HierarchicalClusterer for all the SUC datasets.
For both clustering algorithms and for all the datasets, choose "Classes to cluster evaluation" in the Cluster mode pane. Rember to change the number of clusters according to the number of categories of the dataset at hand.
Create a table to organize your results.

Q2:

- What is the best performance?
- How successful have the clustering algorithms been all in all?
- Looking at each class individually, can you spot particular classes that are consistently well identified by the clustering algorithms?
- Classes that are poorly identified?
- Which classes are mostly confused with each other?
- etc.
- *Provide an interpretation of the clustering results based on the evidences you got.*

VG tasks

Theoretical question:

Q3: Describe k-means' optimization objective in simple words.

Choose the best cluster results you get with **k-means**. To get a concise description of the best clustering produced, we are going to give it to a tree classifier. In the *Visualize cluster assignment* window, select *Save* to output the cluster assignment to a data file. In the data file, replace **Cluster** by **class** in **@attribute Cluster {cluster1, cluster2, cluster3}.** Load this file and apply J48 (disable pruning and keep the parameter M on the default value 2).
Evaluate on the training set and with 10-fold-crossvalidation.

Q4: Do you get a good description of the clusters? Visualize the trees. Is it what you expected? Explain and interpret what you see.

To be submitted

A written report (at least 2 pages) containing the **reasoned** answers to the tasks and questions above and a short section where you summarize your reflections and experience. If you just cut and paste Weka results page into the report without commenting or explaining the whys and wherefores, you might fail the assignment.

Submit the report in **PDF** format to santinim@stp.lingfil.uu.se no later than **8 January 2017, 23:59**. **Please, write this phrase in the subject line of your email: "ML4LT 2016 – Ass3: *your name*".** Attach any additional material that you think is important to fully understand your report. No need to paste in Weka result page in your report in if not needed in your discussion in the report.

Naming conventions

Please, name your pdf report in this way (it will be easier for me to organize and archive them): surname_name_*ass3report*.pdf (ex: *santini_marina_ ass3report.pdf*).

Appendix

SUCs text categories divided into genre, domain and mixed.

	Main Categories	Subcategories	Genre or Domain?
A	**Press, Reportage**		**genre**
		AA. Political AB. Community AC. Financial AD. Cultural AE. Sports AF. Spot News	
B	**Press, Editorials**		**genre**
		BA. Institutional BB. Debate articles	
C	**Press, Reviews**		**genre**
		CA. Books CB. Films CC. Art CD. Theatre CE. Music CF. Artists, shows CG. Radio, TV	
E	**Skills, trades and hobbies**		**domain**
		EA. Hobbies, amusements EB. Society press EC. Occupational and trade union press ED. Religion	
F	**Popular lore**		**domain**
		FA. Humanities FB. Behavioural sciences FC. Social sciences FD. Religion FE. Complementary life styles FF. History FG. Health and medicine FH. Natural science, technology FJ. Politics FK. Culture	
G	**Biographies, essays**		**genre**
		GA. Biographies, memoirs GB. Essays	
H	**Miscellaneous**		***mixed***
		HA. Federal publications HB. Municipal publications HC. Financial reports, business HD. Financial reports, non-profit organisations HE. Internal publications, companies HF. University publications	
J	**Learned and scientific writing**		**genre**
		JA. Humanities JB. Behavioural sciences JC. Social sciences JD. Religion JE. Technology JF. Mathematics JG. Medicine JH. Natural science	
K	**Imaginative prose**		**genre**
		KK. General fiction KL. Mysteries and science fiction KN. Light reading KR. Humour	

--the-end--