Pre-processing, Feature extraction, and Specialization of Nominal Data

Krishna Aryal, Meena Jha, Michael Li

Aim

The main aim of this research is to design a framework that systematically pipelines all the required phases to process nominal data, extract, and select identified features from nominal data based on the dependability of the features, and calculate statistically significant values of aggregated or split attributes.

Objective

The objectives of this research study are as follows:

- To identify the phases required to process nominal data and convert it to a statistically significant value.
- To identify, aggerate, and classify the attributes and find a method to generate a pool of attributes for a given dataset.
- To create relation-based rules with respect to any data domain and apply those to collected attributes.
- To select and categorize attributes by exploring the relations with the targeted attribute in a pool of collected attributes for a given data domain.
- To identify a method to merge the attributes together and calculate aggregated result.

Pre-processing, Feature extraction, and Specialization of Nominal Data

Background

- To use a data efficiently, first we need to understand the nature of it considering the data domain (Perreault Jr & Leigh, 1989).
- Nominal data is different from numerical data which require different processing methods to make it fit for the Artificial Intelligence / Machine Learning (AI/ML) Models (Agarwal et al., 2018; Lakshminarayan, 2013; Perreault Jr & Leigh, 1989).
- Creating bag of words for Text classification, understanding the features based on the distance measurement, summarizing the text based using Natural Language Processing, are preprocessing steps that cleans data and prepare for the aggregation of feature selection and extraction (Agarwal et al., 2018; Kramer et al., 2001; Zhai et al., 2018).
- Algorithms like SVM, Chi-Squared, Gain Ratio, Distance measurement between attributes, Naive bayes are some of the algorithms that helps to select those features after processing it (Agarwal et al., 2018; Li & Gu, 2015; Meng & Xu, 2018).
- Framework brings all steps together that are required to process to process nominal data and give result without any middle intervention (Minsky, 2019).

Pre-processing, Feature extraction, and Specialization of Nominal Data

Proposed Workflow Diagram

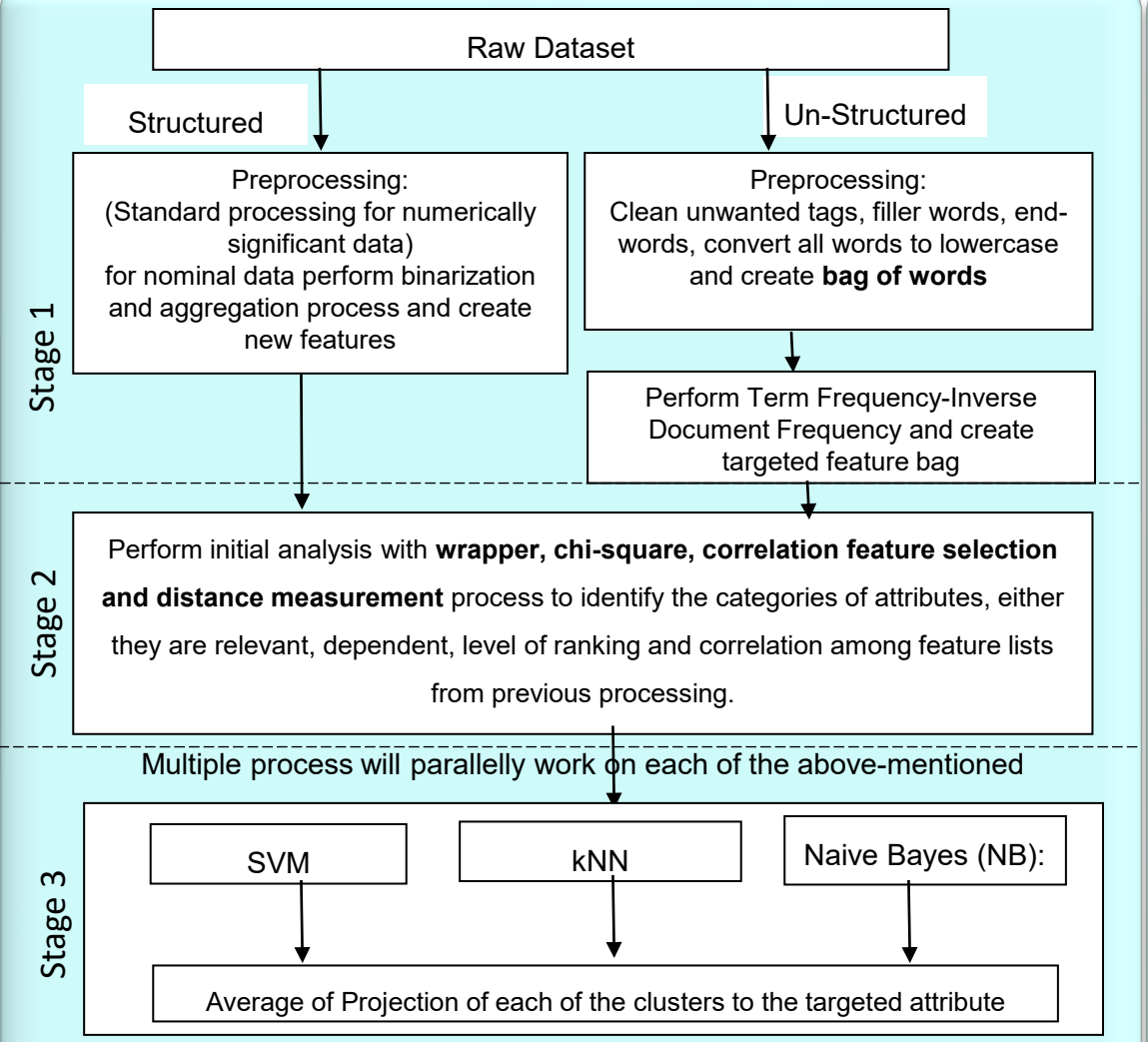

Figure 1: Block Diagram of the proposed system

Methodology to Develop the Architecture

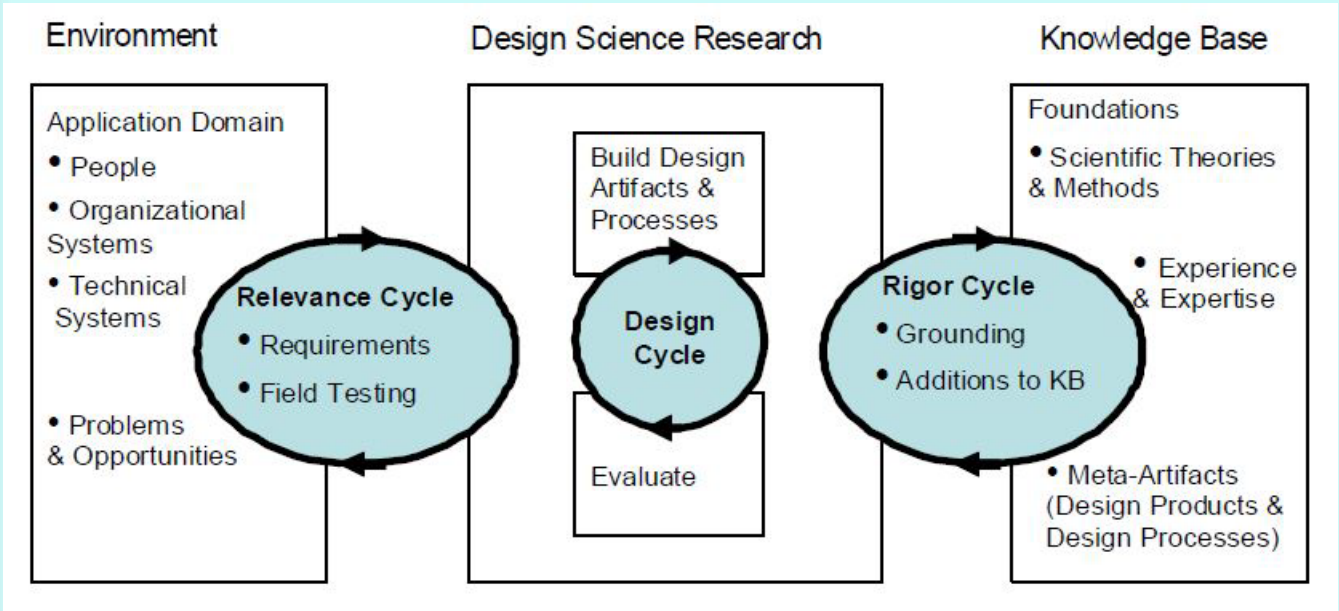

Figure 2: Design Science Research Cycles (Hevner, 2007)

Explanation to use Design Architecture:

1. Sub-systems developed for each stage will act as environment for the following stage. For example, artifact developed for pre-processing will generate the relation to work on the second stage and similarity between stage second and stage third.
2. Knowledge based input for all three stages will be different, depending on the functionality that need to be carried out in the dataset provided.
3. Design Science Research part will implement the inputs and logic for each stage and generate the artifact.

Overall development will create a planned final framework.

Pre-processing, Feature extraction, and Specialization of Nominal Data

Conclusion

Considering the background research and relevant works performed by researchers, there are still some gaps like not addressing aggregated nominal relational attributes with better feature selection algorithms like TF-IDF, chi-squared method and CFS method, not considering distance measurement methods with chi-squared and gain ratio methods and not adding real-world relation of those selected feature with weight-based distance measurement methods. By considering and working on these gaps we can create a new framework that will be efficient and effective to address nominal data. This framework will not be specific to any field rather than it will be based on the nature of provided data and data domain. Future work is to build a prototype and to test it with open data sources to determine the effectiveness.

References

Agarwal, A., Baechle, C., Behara, R., & Zhu, X. (2018). A Natural Language Processing Framework for Assessing Hospital Readmissions for Patients With COPD. *IEEE Journal of Biomedical and Health Informatics*, *22*(2), 588-596. https://doi.org/10.1109/JBHI.2017.2684121

Hevner, A. R. (2007). A three-cycle view of design science research. *Scandinavian journal of information systems*, *19*(2), 4.

Lakshminarayan, N. (2013). Know Your Data Before You Undertake Research. *Journal of Indian Prosthodontic Society*, *13*(3), 384-386. https://doi.org/http://dx.doi.org/10.1007/s13191-013-0300-8

Kramer, S., Lavrač, N., & Flach, P. (2001). Propositionalization Approaches to Relational Data Mining. In S. Džeroski & N. Lavrač (Eds.), *Relational Data Mining,* 262-291. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-04599-2_11.

Li, Z., & Gu, W. (2015). A redundancy-removing feature selection algorithm for nominal data. *PeerJ Computer Science*, *1*, e24.

Meng, F., & Xu, L. (2018). An Improved Native Bayes Classifier for Imbalanced Text Categorization Based on K-Means and Chi-Square Feature Selection. 2018 Eighth International Conference on Instrumentation & Measurement, Computer, Communication and Control (IMCCC).

Minsky, M. (2019). A Framework For Representing Knowledge Frame Conceptions and Text Understanding. In D. Metzing (Ed.),1-25. De Gruyter. https://doi.org/doi:10.1515/9783110858778-003

Perreault Jr, W. D., & Leigh, L. E. (1989). Reliability of nominal data based on qualitative judgments. *Journal of marketing research*, *26*(2), 135-148.

Zhai, Y., Song, W., Liu, X., Liu, L., & Zhao, X. (2018). A Chi-Square Statistics Based Feature Selection Method in Text Classification. 2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS).