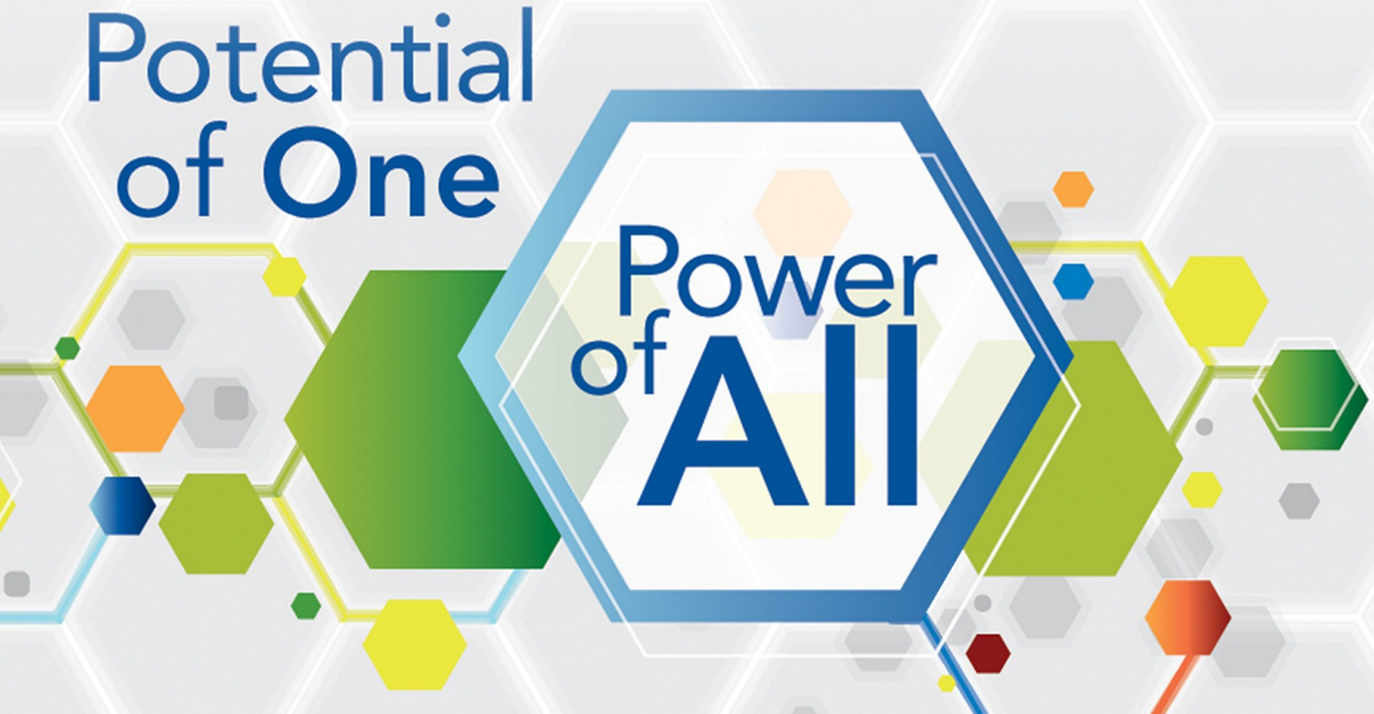
SAS GLOBAL FORUM

An Ensemble Approach for Integrating Intuition and Models

Masoud Charkhabi, Lingyun Zhu
Canadian Imperial Bank of Commerce



An Ensemble Approach for Integrating Intuition and Models

Masoud Charkhabi, Lingyun Zhu
Canadian Imperial Bank of Commerce

Abstract

Finding groups with similar attributes is at the core of knowledge discovery. To this end, Cluster Analysis automatically locates groups of similar observations. Despite successful applications, many practitioners are uncomfortable with the degree of automation in Cluster Analysis, which causes intuitive knowledge to be ignored. This is more true in Text Mining applications since individual words have meaning beyond the dataset. Discovering groups with similar text is extremely insightful however blind applications of clustering algorithms ignore intuition and hence are unable to group similar text categories. The challenge is to integrate the power of clustering algorithms with the knowledge of experts. We demonstrate how SAS/STAT® 9.2 procedures and the SAS Macro Language are used to ensemble the opinion of domain experts with multiple clustering models to arrive at a consensus. The method has been successfully applied to a large dataset with structured attributes and unstructured opinions. The result is the ability to discover observations with similar attributes and opinions by capturing the wisdom of the crowds—whether man or model.

Objective

There are two objectives to this study:

1. Merging unstructured data with structured data and clustering similar observations
2. Merging patterns in data with knowledge outside the data

Method

Ensemble methods are generally geared to two strategies. The first is to learn a joint model and the second is collaborating output for a consensus. The study focuses on the latter. Since ensemble methods involve the collaboration of many models, or opinions, they serve as a scientific was of integrating intuition into models.

Intuition can be injected at two different layers:

1. Clustering Design Decisions: the cluster analysis process includes a variety of design decision prior to implementation. These decisions serve as an appropriate point of entry for intuition. In this study the below phases were used to inject intuition:

- Variable selection
- Standardization and transformation
- Similarity measure
- Algorithm
- Performance criterion

2. Efficient Ensemble Intervention at the Cluster Level: a series of outcomes will be created through the prior process. Intuition can be used to classify these outcomes into three categories.

- Eligible voters
- Privileged voters
- Ineligible voters

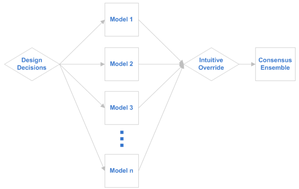


Figure 1. Ensemble process to integrate intuition into models

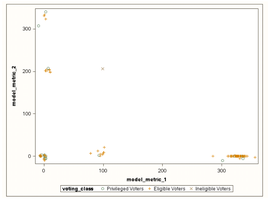


Figure 2. Outcomes according to intuition allows control over ensemble process

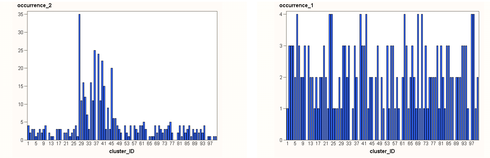


Figure 2. Ensemble methods are effective when a mode exists in the outcomes. This is present in the left hand side figure. In the right hand Side figure an ensemble process will generate a random outcome.

Results

Insights from mixed structured and unstructured data require merging intuition into models. Ensemble methods applied through SAS macros allow an analyst to inject intuition into the modeling process while keeping track of interventions.

Conclusions

Ensemble methods can be a means to inject intuition into models with reproducible results. Often intuition is considered intangible and difficult to explain. This study Ensembles allow analysts to inject tuition into various phases of the process without sacrifices rigidity. The injection of intuition into models is even more essential in text mining applications.

References

- J. Gao, W. Fan and J. Han. On the Power of Ensemble: Supervised and Unsupervised Methods Reconciled. A Tutorial on SIAM Data Mining Conference (SDM), Columbus, OH, May 1 2010.
- V. Ilango, R. Subramanian and V. Vasudevan. Cluster Analysis Research Design model, problems, issues, challenges, trends and tools. International Journal on Computer Science and Engineering. volume:3 issue:8. pp. 2926-2934. 2011.
- D. Parra and X. Amatriain. Walk the talk - analyzing the relation between implicit and explicit feedback for preference elicitation. Proceedings, In User Modeling, Adaption and Personalization - 19th International Conference UMAP. Girona, Spain. pages 255-268. 2011.
- S. Vega-Pons and J. Ruiz-Shulcloper. A Survey of Clustering Ensemble Algorithms. International Journal of Pattern Recognition and Artificial Intelligence. Vol. 25, No. 3 (2011) pp. 337-372.

SAS® GLOBAL FORUM

Washington, D.C.

March 23–26, 2014