Cluster Ensembles, Majority Vote, Voter Eligibility and Privileged Voters

Masoud Charkhabi, Tarundeep Dhot, and Shirin A. Mojarad

***Abstract*—Grouping items facilitates ideation. Although Cluster Analysis has become a classical technique for grouping in science and engineering, to the best of our knowledge it's use remains limited in business. In this application paper we use cluster ensembles to address three barriers to wide scale adoption in the banking industry. The aforementioned challenges are: consistency of results, knowledge beyond the data and grouping with multiple objectives. Contributions of this study include guidance on dealing with the lack of meaningful cluster labels (in the case of ensembles), bimodal cluster distributions and incorporating expert intuition into the clustering process. This application has delivered unobvious insight into a high-dimensional dataset to audiences with diverse backgrounds.**

***Index Terms*—Clustering, cluster ensembles, majority vote.**

I. Introduction

"Facts come from negotiation between different parties."

— Paul Feyerabend

Retail banks manage three forms of physical assets; bank branches, Automated Banking Machines (ABMs) and mobile advice centers. Large amounts of data are accumulated from the sensors and systems attached to these assets. Furthermore, the geospatial qualities of these assets make public geographical data sources applicable. The result is a high-dimensional dataset of items. Grouping the assets based on similar attributes facilitates innovative strategies for placement and investment optimization. The traditional approach to grouping these assets is a hierarchical partitioning based on univariate distributions of key attributes. For example, grouping based on the population of the Census Metropolitan Area (CMA) that the asset resides in and the revenue generating ability of the asset. These approaches are extremely sensitive to the sequence of items chosen for partitioning. Cluster analysis is becoming more mainstream in the financial services industry for segmenting customers; yet barriers remain to wide scale adoption. The randomness embedded in cluster analysis assigns items to different groups when slight modifications are made to the cluster generating algorithm or data transformation process. This is more concerning when clustering procedures are placed in sequence. By this we mean the output of one process is the input to the next. In addition, grouping at various discrete time intervals is often insightful; this involves running the clustering algorithm multiple times, often with slight modifications [1].

We empathize with practitioners that are uncomfortable with the level of automation of clustering algorithms. Intuitive outliers and deterministic partitions are often ignored in blind applications. There is a need to incorporate expert intuition that resides outside of the dataset or learning algorithm. Grouping is rarely without an objective [2]. As asset management involves continuously changing objectives, data must be grouped in different ways. Often this is done in sequence where the output of a clustering procedure is the input to another. In the case of sequential clustering, errors may accumulate uncontrollably. The three described challenges; consistency of results, knowledge beyond the data and grouping with multiple objectives can be addressed with cluster ensembles. We defer the discussion of general challenges in clustering to [3].

Cluster ensemble methods consist of two stages: generating clusters and calibrating the results to arrive at a consensus [4]. The calibration greatly stabilizes the process and hence addresses the challenge of consistency in results. Cluster ensembles used in sequence tend to relax concerns of accumulating error since the individual steps are calibrated to reduce error. And knowledge beyond the data is incorporated by interfering in the consensus stage.

Establishing an analog is useful in explaining how expert intuition is injected into the process. Assume an election process where the output of each cluster algorithm (in an ensemble of clustering runs) serves as a vote. A domain expert can define a set of eligible voters before the consensus process (in the ensemble case) and hence, incorporate the desired bias. In certain cases, the eligible voter strategy is not strict enough and unquestionable knowledge beyond the data has to be inserted. In this case, the domain expert intervenes as a privileged voter and overrides all other votes. "Voter Eligibility" and "Privileged Voters" allow the control that is demanded for comfortable wide-scale adoption. Voters from the algorithm and the domain expert often have competing opinions. Our experiment shows that the calibration of opinion has advantages over any single opinion. We also expose weaknesses in predecessor stages of clustering (in the ensemble case) as well as undesired expert bias.

Contributions of this study are the findings that come from the application of the cluster ensemble method to address the three described challenges. Three main findings have transpired. The first stems from the fact that the output labels generated by various cluster runs are difficult to compare. This challenge is irrelevant in supervised learning where

Manuscript received September 19, 2013; revised November 30, 2013. This work was supported in part by the Canadian Imperial Bank of Commerce. The views expressed in this paper are solely those of the authors.

Masoud Charkhabi, Tarundeep Dhot, and Shirin A. Mojarad are with the Advanced Analytics division of the Canadian Imperial Bank of Commerce. Commerce Court, Toronto, Canada (e-mail: masoud.charkhabi@cibc.com, tarundeep.dhot@cibc.com, shirin.mojarad@cibc.com).

DOI: 10.7763/IJMLC.2014.V4.424

successful ensemble methods motivated the use of ensembles in clustering [5]-[7]. The second finding is that the ensemble process exposes weakness in preceding assumptions. Details are discussed in Section II-B. The third finding is that the practice of establishing eligible voters and privileged voters improves the quality and the interpretation of clusters. Thus, using cluster ensembles addresses the three main barriers to wide scale adoption. Aligned with these areas of concern are the three core findings and experimental results in support of them.

II. Experiment

The objective of this exercise was to facilitate idea generation by discovering natural groups in the dataset of assets. The cluster ensemble method is used to address challenges of consistency in results, knowledge beyond the data and multi-objective grouping. In this section, experiments are described that evaluate the success of the ensemble method in addressing these challenges. The focus is on the consensus stage of the ensemble process. Three stages of clustering are performed on a dataset of bank physical assets. The underlying data can be described as high-dimensional with diverse attributes (nominal, categorical, binary and intervals scaled). The details of the data are proprietary and present little value to the paper.

A. Data Preparation

Some level of manual partitioning of attributes and observations is required as a pre-processing step. In some ways, this initial treatment incorporates knowledge beyond the data and facilitates multi-objective clustering. The trade-off is that bias is introduced. Standardization and imputation are necessary tasks in real world data mining applications. Their importance in cluster analysis is highlighted in [3]. These tasks play a more strategic role in partitioning when using cluster ensemble methods. This is similar to the role that these two data preparation steps play in supervised learning in that they are key tuning parameters for improvement. Imputation is regarded as craftsman's work that should be approached with caution [8]. We observed that a consequence of imputation in clustering is that clusters with better statistics are created artificially. This is due to the objective of imputation which is to replace the unknown with similar values from the dataset. In some ways, the process creates copies of data and hence better natural clusters. Without imputation, significant data can be sacrificed and the interpretation of the results becomes difficult.

B. Cluster Generation

We assume the number of clusters has been determined a priori. Various clusters are generated by varying the random seed and clustering algorithm. In this experiment, two clustering algorithms are used: k-means and kernel k-means clustering. Kernel k-means clustering is a generalized form of standard k-means clustering algorithm, proposed to identify non-linearly separable clusters by implicitly mapping inputs to a higher dimensional space which can be scaled to large data sets [9].

K-means and kernel k-means clustering algorithms are run with 100 different random seeds each. The result is a dataset with 200 cluster labels for each observation. Evaluation of the cluster validity indexes [10], [11] (Cubic Clustering Criterion, F-statistic, within cluster standard deviation and R-squared) show that many of the generated clusters are poor performers (see Fig. 1). The quality of clusters will suffer if all generations are passed to the consensus stage. A cut-off for performance is established that limits the generations to 100. This filter is statistically motivated and not based on knowledge beyond the data. Fig. 1 shows the 100 alternatives created and their corresponding cluster validity indexes. The concentration of cluster validity indexes in the three areas of the graph highlights the value of generating many alternatives.

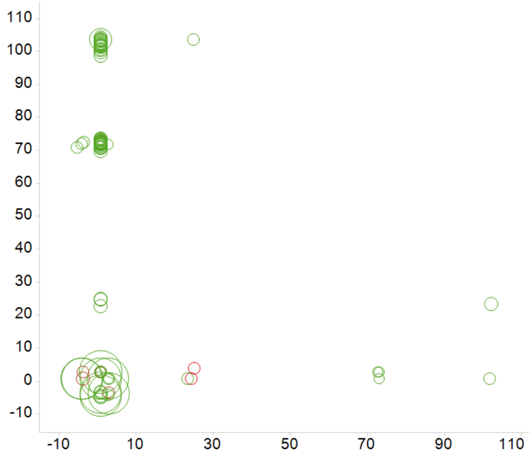

Fig. 1. Cluster validity indexes for 100 alternative cluster procedures. CCC on the x-axis, pseudo F-statistic on the y-axis, Within Cluster Standard Deviation (WCSTD) as the red-to-green increasing gradient and R-squared as the bubble size.

```
        Generate many cluster labels
        Select a random generation A
              Repeat (for 1 to C):
                 Locate nearest neighbor cluster in A
        to all generations
               Label the nearest neighbor
                 Remove the nearest neighbor from
        the pool
Repeat
```

Fig. 2. Program used to build comparable cluster labels using an iterative nearest neighbor approach without replacement where *C* is the number of clusters and *A* is a random cluster generation.

C. Establishing Consensus

We refer back to the analog described in the introduction. Consistency in results is measured by comparing the distribution of the majority vote (from the ensemble) to a random individual vote (from one clustering procedure). Two challenges prevent a straight forward comparison between voters. The first is the fact that labels from clustering algorithms are meaningless in isolation. The comparison between labels must leverage the description of the cluster. This description is captured through the cluster mean attributes. The approach used in this study was to compare the cluster mean attributes from various cluster generations to land at a nearest neighbor cluster. The program is described in Fig. 2 where *C* is the number of clusters and *A* is a random cluster generation. The process must avoid random selection with substitution, to ensure that the total number of clusters remains stable. The program is similar to a hierarchical clustering process but customized to this application. The conclusion is that cluster labels in different generations are now consistent and comparable.

The second challenge arises in the consensus stage. When the distribution of votes is near uniform, the consensus is merely a coin tossing exercise. This phenomenon is measured through the level of cardinality in the distribution of clusters, the frequency of the mode and the count of known anomalies. Purity, inverse purity and the F-statistic are measures that are commonly used to capture the quality of ensembles [12]. These metrics capture essentially the same information as the ones we used. The performance that is desired is low cardinality, high frequency of mode and low count of anomalies. Intuitively, the baseline for the initial run is randomness. Once more experience is gained, the ensembles can be compared. Table I shows the performance of the first three ensemble methods with non-trivial gains above randomness. The second and third ensembles incorporate more experience by generating more clusters which improves the performance slightly. With many generations, the exact statistic used becomes less important [11]. These statistics are appropriate to measure consistency advantages of the ensemble method.

TABLE I: ENSEMBLE CONSISTENCY

Scenario	Ensemble Performance Indexes		
	Average Cardinality	Average Mode Frequency	Anomalies
Ensemble 1	12.87	7.14	1.3%
Ensemble 2	12.11	7.46	1.3%
Ensemble 3	11.49	7.95	1.4%

D. *Incorporating Domain Expertise*

The degree to which knowledge beyond the data has been incorporated into the process is qualitative. What can be observed is the comfort level of the user, and the transparency in which bias is injected. The user base of this study is too small at this stage to conduct a meaningful survey. The transparency of the approach used in this study is worth describing.

Domain experts offer an opinion. Between the extreme of full reliance on this opinion and blind acceptance of a statistical learning algorithm lies a spectrum of alternatives. The ensemble approach, combined with the strategy of voter eligibility and privileged voters presents a wide range of options across this spectrum. In Section II-B we described the process of building a dataset of alternative cluster labels for each observation, creating comparable labels and limiting cluster outcomes with cluster validity indexes. We capture the opinion of the domain expert in two stages. In the first stage, alternatives are limited to eligible voters. The domain expert evaluates random clusters and observations from each generation and decides which generations should proceed as eligible voters. In the second stage, observations of particular interest are evaluated where overwhelming evidence suggests a manual over-ride. This decision creates the "privileged voter" (the voter that heavily influences the outcome for reasons beyond the data). The expert injects a high-frequency of votes for these "privileged voters" to reflect this strong opinion. The mode of all eligible and privileged voters becomes the final label.

Fig. 3 shows the distribution of alternatives for four observations. The bottom two distributions are examples of multi-modal scenarios and the top two distributions are examples of good consensus. The privileged voter override is very useful in scenarios where the mode is unconvincing. In Fig. 4, the domain expert has highlighted eligible voters in blue.

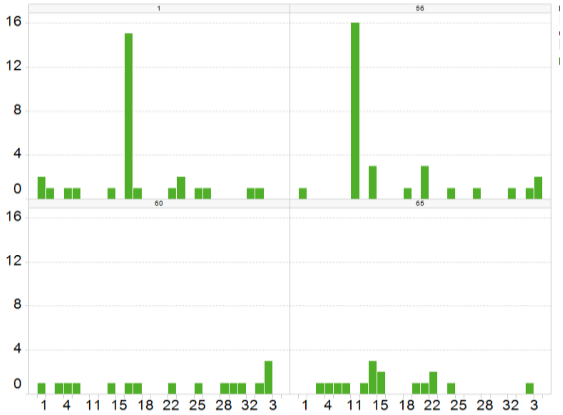

Fig. 3. Distribution of cluster labels for four observations. The bottom two distributions are examples of multi-modal scenarios and the top two distributions are examples of good consensus (at the peaks).

.

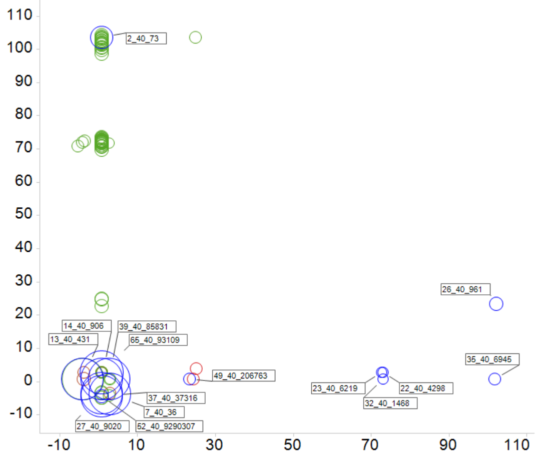

Fig. 4. Eligible voters are highlighted by the domain expert. CCC on the x-axis, pseudo F-statistic on the y-axis, Within Cluster Standard Deviation (WCSTD) as the red to green increasing gradient and R-squared as the bubble size.

III. RELATED AND FUTURE WORK

Ref. [11] presents a modern and comprehensive survey of cluster ensemble methods and [5] shares experiences from simulation studies. General discussions on challenges in cluster analysis are reviewed in [13]. It is useful to learn from ensemble methods in supervised learning, specifically on nearest neighbor classifiers due to their similarity to clustering [14], [15]. The calibration of expert opinion and clustering algorithms is similar to the concept of integrating explicit and implicit feedback. The work of the authors of [7] inspired this approach. Clustering with bias is not only acceptable but necessary in many applications [9].

Future work will focus on more rigorous evaluation measures for ensemble evaluation and an optimization function for the final consensus. In this phase of the study, control and clarity of objects co-occurrence was preferred over the theoretical soundness of median partition [11]. We felt good practice would be to lead with a controlled and transparent approach and iteratively increase rigor and automation.

IV. Discussion and Conclusions

Three challenges to the wide-scale application of cluster analysis were highlighted: inconsistent cluster results, the importance of knowledge beyond the data and clustering with many objectives. The advantages of cluster ensemble methods were described as intuitive solutions to these concerns. The experiment described in this study provides support for this intuition. Improvements over single-generation clustering are evident in cardinality related measures and visualizations. Three challenges are faced when applying cluster ensembles: lack of labels, the need for involving expert intuition and concerns of amplifying error in sequential multi-clustering. A distance program creates comparable labels. Filtering eligible voters and defining privileged voters incorporates expert opinion in a transparent fashion. In a real world application, this serves as a healthy dose of bias in the initial setup, to increase user comfort. The intent is to gradually relax bias. Multiple objectives can be achieved by partitioning the data selectively prior to clustering (selecting different observations and variables from a larger comprehensive dataset). The calibration feature of ensemble methods reduces the concern that the error will become uncontrollable. Consistency is the main issue in multi-layer clustering. Calibrating across cluster generations and domain experts can be seen as a form of human-computer cooperation—advantages of which are prevalent in data mining studies [7].

References

[1] W. Claster, S. Shanmuganathan, and P. Sallis, "Wine tasting and a novel approach to cluster analysis," in *Proc. 4th Asia International Conference on Mathematical/Analytical Modelling and Computer Simulation*, Kota Kinabalu, Malaysia, 26-28 May 2010, pp. 152-157.

[2] S. X. Yu and J. Shi, "Grouping with bias," *Advances in Neural Information Processing Systems*, vol. 14, 2001, pp. 1327-1334.

[3] 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*, vol. 3, no. 8, pp. 2926-2934, 2011.

[4] 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, pp. 337-372, 2011.

[5] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, New York, NY: Springer, 2008.

[6] J. Ghosh and A. Acharya, "Cluster ensembles," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 1, no. 4, pp. 305-315, 2011.

[7] T. K. Paul, Y. Hasegawa, and H. Iba, "Classification of gene expression data by majority voting genetic programming classifier," *IEEE Congress on Evolutionary Computation*, Vancouver, Canada. 2006.

[8] E. Rancourt, "Estimation with nearest neighbour imputation at statistics canada," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 1999.

[9] T. H. Sarma, P. Viswanath, and B. E. Reddy, "A fast approximate kernel k-means clustering method for large data sets," *Recent Advances in Intelligent Computational Systems (RAICS)*, pp. 545-550, 22-24 Sept. 2011.

[10] SAS Institute Inc., SAS Technical Report A-108, *Cubic Clustering Criterion*, Cary, NC: SAS Institute Inc., pp. 56, 1983.

[11] Y. Yan, R. Rosales, G. Fung, and J. G. Dy, "Active Learning from Crowds," in *Proc. the 28th International Conference on Machine Learning*, Bellevue, WA, USA, 2011.

[12] X. Z. Fern and C. E. Brodley, "Solving Cluster Ensemble Problems by Bipartite Graph Partitioning," in *Proc. the 21st International Conference on Marchine Learning*, Banff, Canada, 2004.

[13] G. James, "Majority vote classifiers: theory and applications," Ph.D. dissertation, Dept. Statistics, Stanford Univ., 1998.

[14] L. I. Kuncheva, C. J. Whitaker, C. A. Shipp, and R. P. W. Duin. (2003). Limits on the majority vote accuracy in classifier fusion. *Pattern Analysis and Applications*. [Online]. 6. pp. 22-31. Available: http://pages.bangor.ac.uk/~mas00a/papers/lkpaa.pdf

[15] D. Parra and X. Amatriain, "Walk the talk - analyzing the relation between implicit and explicit feedback for preference elicitation," in *Proc. 19th International Conference on User Modeling, Adaption and Personalization*, Girona, Spain, 2011, pp. 255-268.

Masoud Charkhabi is the director of Advanced Analytics at the Canadian Imperial Bank of Commerce (CIBC). The Advanced Analytics unit was established two years ago with a mandate to facilitate knowledge discovery and decision support using statistical learning and data mining methods. The scope of the group spans across CIBC's vast structured and unstructured data sources. Previously, Masoud held consulting and management roles in the decision sciences, technology and operations divisions of CIBC. He holds an academic degree in mechanical engineering and management.

Tarundeep Dhot is a senior consultant in the Advanced Analytics unit at the Canadian Imperial Bank of Commerce. Tarun has held consulting and modeling roles in the decision sciences, fraud strategy and operations divisions of CIBC. He holds academic degrees in computer science (Master) and electrical engineering (Undergraduate).

Shirin A. Mojarad is a senior analytics specialist in the Advanced Analytics unit at the Canadian Imperial Bank of Commerce. Previously, Shirin has worked in the industry as a data mining consultant. She obtained her Ph.D. in electrical engineering from Newcastle University, U.K., where she specialized in predictive modeling and neural.