Dynamic Classification Unit (DCU)
for Online Segmentation of Big Data via Small Data Buffers

Anna Khalemsky & Roy Gelbard

Information Systems Program, Bar-Ilan University, Israel

Introduction

In various classification processes, new cases are classified according to a model that was built on the basis of past cases. As long as the new cases are "similar enough" to past cases, the classification runs normally. However, when a new case is substantially different from the known cases, a reexamination is required. Since in big-data environment it is not possible to reexamine *all* past data, the current research presents a model in which small groups of selected cases are stored in **data buffers**, and an **incremental dynamic classifier** dynamically updates the segmentation sets (classification categories). The model enables automatic dynamic decision-making in real time, and provides an effective tool for dynamic big data environment.

Background

"Since the amount of data being processed is large, it is important for the mining algorithms to be very computationally efficient. Recently many important applications have created the need of incremental mining“ (Shah, Chauhan & Bhanderi, 2012). The main advantage of algorithms that focus on incremental dynamic analysis is the significant savings in resource utilization and the speed at which they absorb and process new information that flows dynamically, since they permit use of small part of the data (Song, Meng, Wang, O'Grady & O'Hare, 2009). Data mining has a large toolbox of models and algorithms that are adapted to a wide range of problems. Adaptation of these tools must take into account both objective and subjective aspects of similarity indices in order to have a better fit to organizational needs (Gelbard, Goldman & Spiegler, 2007).

Research method

1) **Preliminary evaluation of research rationale**. To evaluate the ability of clustering algorithm (without labeled data) to yield good predicting results, we symbolized “cluster” to a “path” (from root to leaf) in a decision tree, which presents classification rules. In this way we check the assumption that a clustering algorithm is almost as good as a decision tree in terms of the accuracy of the results it provides.
2) **Evaluation of the proposed model**. The model was evaluated using different configurations of parameters (e.g., buffer size, initial number of clusters, threshold level, and normalization/standardization technique). Different threshold levels of RMSE enabled us to detect significantly "different" cases which can indicate the formation of a new segment.
3) **Data**. The proposed model was evaluated using two datasets: (1) "**Lev**“, a dataset containing 1,000 anonymous ordinal evaluations of lecturers, donated by Prof. Ben David; and (2) "**Occupancy Detection**“, a dataset containing 20,560 records on ground truth occupancy, based on five numerical variables, downloaded from the UCI machine learning repository.

Model Architecture & Logic

Real-time data flows from the "***Sensor***" (Figure-1) to the dynamic classification unit (DCU). The DCU performs a decision process based on a dynamic segmentation scheme (Figure-2). The classifier incrementally updates the populations of the relevant “***Case buffers***”, based on the threshold of the segmentation error parameter and the segmentation quality criteria. The mechanism that manages the cases stored in each buffer can use different policies, such as “FIFO” (First-In First-Out), and “Archetypes” (a policy that stores the extreme cases of each segment - the "outliers" that are still classified to this segment) .

The term "***Sensor***" denoted the "funnel" through which the data stream flows. Thus, a sensor can be a physical object, as well as a logical handshake. Each DCU is a remote autonomous “sensing” system that can communicate with multiple additional DCUs, as well as with a “***Central Controller***”.

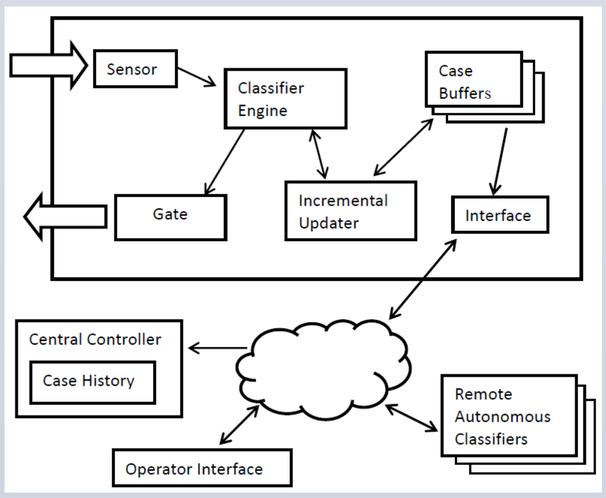


Figure-1: DCU Architecture

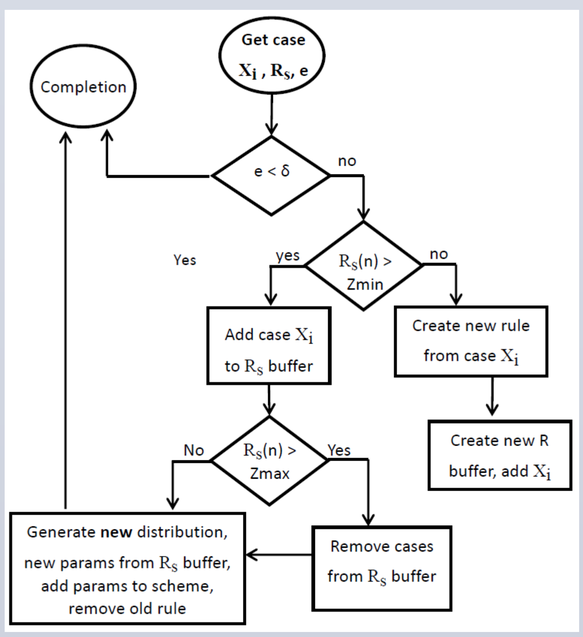


Figure-2: Dynamic Segmentation Logic

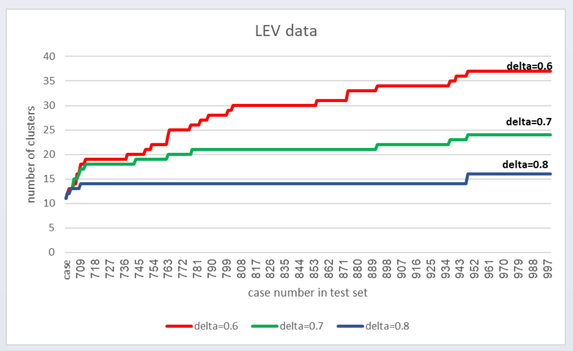
Results

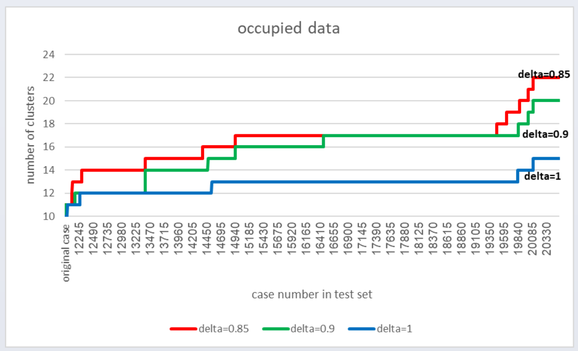
1) Interpreting and evaluating clustering results via analogy to decision trees

Data (normalized/standardized)	Tree algorithm	Tree accuracy	Number of groups	Cluster accuracy
LEV - normalized	J48	60.4%	65	49.5%
	RepTree	60.1%	30	52.3%
	Random Tree	62.8%	90	47.6%
Occupied – normalized	J48	99.2%	42	94.44%
	RepTree	99.01%	40	94.43%
	Random Tree	99.2%	186	93.53%
Occupied - standardized	J48	99.22%	42	86.43%
	RepTree	99.09%	40	86.37%
	Random Tree	99.16%	183	87.04%

The clustering technique without labeled data succeeds in achieving results that are similar to those of classification tree, which can be used only with labeled data.

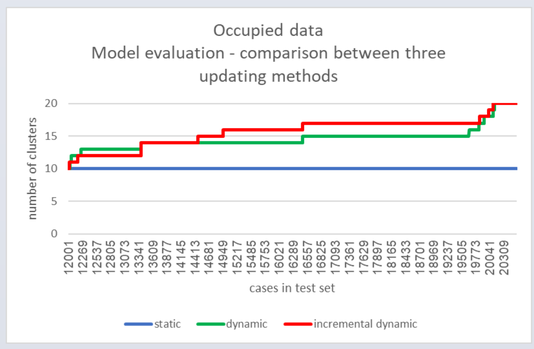
2) Convergence of dynamic segmentation processes using small memory buffers





The two datasets detect the convergence of segmentation processes and, despite the partial data, the model succeeds to organize the cluster set very well.

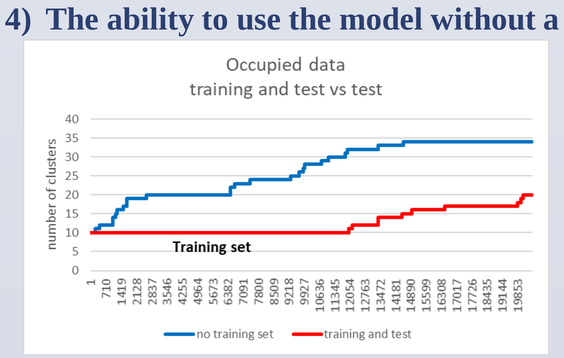
3) Model efficiency



	Static state	Dynamic state Non-Restricted Buffer	Incremental dynamic State Buffer=100
N. of clusters at end point	10	20	20
Average RMSE	0.5167	0.4366	0.4817
Std.dev. (RMSE)	0.2887	0.1884	0.2118
Running Time	108 sec	138 sec	121 sec

Advantage (in running time) of incremental dynamic state over dynamic state: 14%

4) The ability to use the model without a training set



The model succeeds in achieving convergence without a ***training*** phase: most of “additional” segments are outliers, that remain nonfunctional and do not become a “***new trend***”. The mechanism learns and creates rules along the first phase and yields good representative segments.

Conclusions

The novelty of this real-time model lies in the fact that the entire process is based on the use of limited memory buffers. In addition, each DCU, which is a remote autonomous agent, can communicate with multiple additional DCUs to support a distributed environment, regardless the existence of a central controller.

The study illustrates the computational advantage of the incremental dynamic approach over the static and dynamic approaches. Additionally the model can function without relying on any previous information, such as a training set or labeled data.

Further Research

1) Synchronization between multiple agents (DCUs) for optimal operation.
2) Differentiation between new trend detection and exceptional cases (outliers).

References

1. Gelbard R., Goldman O. & Spieger I, (2007). Investigating diversity of clustering methods: An empirical comparison, Data & Knowledge Engineering, 63(1), 155–166.
2. Song, Y. C., Meng, H. D., Wang, S. L., O'Grady, M., & O'Hare, G. (2009). Dynamic and incremental clustering based on density reachable. In INC, IMS and IDC, 2009. NCM'09. (pp. 1307-1310). IEEE.
3. Shah, S., Chauhan, N. C., & Bhanderi, S. (2012). Incremental mining of association rules: a survey. International Journal of Computer Science and Information Technologies, 3(3), 4071-4074.

Contact

anna.khalemsky@gmail.com Roy.Gelbard@biu.ac.il

Patent Application : US 2017/0185866 A1