Incremental Learning

Syeda Badrunnisa Begum
Department of Computer Science & Engineering
RYMCE
Bellary, India

Asst Prof. Suresh
Department of Computer Science and Engg
RYMCE
Bellary, India

***Abstract*— As the organizational data is increasing, the one-shot processing of data for the extraction of knowledge has almost become impossible. Literature has identified the challenges of the growing volumes of data and the knowledge extraction from it; and has coined the extraction of knowledge from such large volumes of data as a data mining [1, 2, 3, 4]. One of the possibilities to extract knowledge from such large volumes of data seems to be incremental learning and therefore incremental learning has become necessary in data mining.**

Incremental learning is a process of deriving the knowledge in phased manner. When there is huge amount of data, then it cannot be retrieved in one-shot. Hence incremental learning gathers the data packets in incremental mode.

OBJECTIVES OF OUR PROJECT IS

- Since learning in an unsupervised scenario is more demanding[5], we would like to employ clustering as a learning mechanism for realizing incremental learning. In particular, we would like to explore the suitability of few of the different clustering algorithms for Incremental learning.
- Heterogeneity of data is posing lot of challenges for learning. We would like to explore mechanisms to convert the heterogeneous data to homogeneous before applying learning mechanisms over it. Since histograms have the generic ability to characterize any type of data [6], we would like to utilize histograms to store the statistical details of the learnt concepts on the way to Incremental learning.

Different combinations of processing the data may not always result in the same output at the advanced stages of incremental learning. Literature has coined it as "order effects" of incremental learning. We would also like to explore mechanisms to minimize or avoid such order effects.

INTRODUCTION

The organizations maintain their organizational data in databases, most of these organizations have online services and they also have to maintain the customer/end-user information. As the organization grows, the databases also grow and contains huge amount of data. Example online learning programs, online shopping, etc.

To obtain the information regarding particular customer/student from large database is possible with the help of data mining. Data mining is a process of retrieving only the required data from the database/data warehouse.

As the organizational data is increasing the one-shot data mining has become impossible. Therefore incremental learning has become necessary in data mining.

1) Incremental Learning can done in 3 forms depending on, how much data, from which prior knowledge was generated [7, 8, 9, 10].

(i) *Zero memory learning* – in which none of the earlier training examples are retained. It is more economical.

(ii) *Full memory learning* – in which all the past training examples are retained, it is likely to result in more accurate descriptions.

(iii) *Partial memory learning* – retains some of the past training examples that are most likely to be of use later on.

2) Different Clustering Algorithms that can be applied to Incremental Learning

Clustering is process of grouping the similar objects in a cluster; these objects are dissimilar to the objects in other clusters. Clustering algorithms can be categorized as Partitioning, Hierarchical, Density-based, Grid-based, Model-based, Constraint-based, and so on. Among these algorithms, density based algorithms have found their suitability in incremental learning [11]. Some of these algorithms are,

- DBSCAN (Density Based Spatial Clustering of Applications with Noise), is a density based clustering technique and has incremental learning ability. It is traced that frequent re-clustering may be needed in a constantly changing database.

- BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies), uses a hierarchical data structure called CF-tree, to incrementally and dynamically cluster the incoming data points. The algorithm BIRCH is a CF tree based multiphase clustering method.

3) Order Effect on Incremental Learning

While learning incrementally from the temporal data, we cannot afford the option of processing the incoming chunks of data in different orders as generation of knowledge at the earliest is the priority. However, if the received data chunks arrive from different sources to a centralized computer, then different options of processing the chunks become possible and there is need to sequence the chunks for processing. The out of order processing of these chunks may affect the overall statistics (knowledge).

In general from an i^{th} data packet $[D]_I$ i^{th} knowledge set $[K]_i$ can be derived. Sooner the $[K]_i$ is available, it has to be merged with the knowledge updated until the end of $(i-1)^{th}$ stage. Then $[Knowledge]_i \leftarrow f([Knowledge]_{i-1}, [K]_i)$

Where

$[Knowledge]_{i-1} \leftarrow f([Knowledge]_{i-2}, [K]_{i-1})$.

If the data is retrieved from multiple sources then order affect problem has to be modeled.

NEED FOR INCREMENTAL LEARNING

- Incremental Learning is ubiquitous in learning.
- One-shot Learning may not be doable at all circumstances.
- This necessitates the need for the development of Incremental Learning System.

One-shot learning

- Learning that takes place over time rather than as a one-shot experience – Christoppe G.C [8]
- A learner L is incremental if

➔ L inputs one training experience at a time.

➔ Does not reprocess any previous experience at a time.

➔ Retains only one knowledge structure in memory – Langley.P[28]

Drawbacks of existing incremental learning systems/algorithms:

- Most of the existing Incremental Algorithms undoubtedly process one experience at a time, but are involved in the reprocessing of the previous experiences and maintains more than one knowledge structure in memory.
- There is lot of inconsistency about the nature of Incremental Learning and very limited efforts were found in analyzing or eliminating the effects of training orders.

Benefits of Incremental Learning:

1. Computational ease and efficacy.
2. Knowledge can be derived in phased manner, so the gradual changes in the knowledge generation process can be visualized.
3. When it is required to learn from unbounded stream of data , if there is no significant change in generated knowledge, the learning process can be terminated.

The learning mechanism employed in this project is CLUSTERING .

- Clustering is a process of partitioning of data set into subsets(clusters) so that data in each subset ideally share some common characteristics.
- Clustering Algorithms differ significantly in their notion of what constitutes a cluster and how to efficiently find them.
- Popular notions include

a. Dense areas of the data space

b. Groups with low distances among the cluster members and so on.

- *Density Based Spatial Clustering of Applications with noise (DBSCAN).*
- *DBSCAN is a density based clustering algorithm because it finds a number of clusters starting from the estimated density distribution of corresponding nodes.*
- *DBSCAN requires two parameters*

a) 1. $\mathcal{E}$ (eps).

b) 2. The minimum number of points required to form a cluster(minpts).

Knowledge parameters used

- Histograms have the generic ability to characterize any type of data.
- Further, *frequency distribution* of elements is conveniently recorded in Histograms.
- Memory required by histogram can be reduced to just two variables irrespective of the number of bins by transforming the histograms to regression line [Lal04, Pra06].
- So, Histogram based regression line can act as a very good knowledge representative of a cluster.
- A group of statistical variables comprising the Number of Elements, Mean, Standard Deviation and Regression Line is used to represent the knowledge of each feature of a cluster and is termed as Knowledge Packet (KP).
- A typical knowledge packet can look like:

TABLE.I Knowledge-Packet

$[B]_i$	No. of elements	f_1			f_2			...	f_n		
$[Cl_1]$	n_1	$^i\mu_1^1$	$^i\sigma_1^1$	$^i L_1^1$	$^i\mu_1^2$	$^i\sigma_1^2$	$^i L_1^2$	...	$^i\mu_1^n$	$^i\sigma_1^n$	$^i L_1^n$
$[Cl_2]$	n_2	$^i\mu_2^1$	$^i\sigma_2^1$	$^i L_2^1$	$^i\mu_2^2$	$^i\sigma_2^2$	$^i L_2^2$	...	$^i\mu_2^n$	$^i\sigma_2^n$	$^i L_2^n$
...	...	..	...	...	..	...	...	...	..	...	...
$[Cl_k]$	n_k	$^i\mu_k^1$	$^i\sigma_k^1$	$^i L_k^1$	$^i\mu_k^2$	$^i\sigma_k^2$	$^i L_k^2$	...	$^i\mu_k^n$	$^i\sigma_k^n$	$^i L_k^n$

Where, μ – Mean; σ – Standard deviation;
L – Regression Line in terms of slope and intercept; Cl – Cluster;

REFERENCES

1. Arun K.P (2005) "Data Mining Techniques", Universities Press ISBN: 81 7371 3804; First Edition; Eighth Impression 2005
2. Han & Kamber, "Data Mining – Concepts and Techniques", Second Edition; Elsevier 2006; ISBN -10: 81-312-0535-5
3. http://www.cs.umn.edu/~han/dmclass/cluster_survey_10_02_00.pdf (Introduction to Cluster Analysis for Data Mining)
4. U, Piatestsky S & Smyth M.A, Uthurusamy R. "Advances in Knowledge Discovery and Data mining", A text Book by AAAI/MIT Press 1996
5. Anil K Jain, M.N Murthy and P.J Flynn, "Data Clustering: A Review", ACM Computing surveys, vol 31, No. 3, September 1999
6. Pradeep Kumar R, "Wavelets for Knowledge Mining in Multi-Dimensional Generic Databases", PhD Thesis of the University of Mysore, 2006.
7. Maloof A.M & Michalski R.S, "A partial memory Incremental Learning Methodology And Its Application To Computer Intrusion Detection", Machine Learning and Inference Laboratory, George Mason University, 1995 (www.mli.gmu.edu/michalski)
8. Maloof A.M & Michalski R.S, "Learning Evolving Concepts Using Partial Memory Approach", 1995 http://www.mli.gmu.edu/papers/91-95/maloof.aaai95.pdf
9. Maloof A.M & Michalski R.S, "Selecting Examples for Partial Memory Learning", Machine Learning, 1-28, Kluwer Academic Publishers, Boston, 1999.
10. Michalski R.S & Larson J.B, "Incremental Generation of VL1 Hypotheses: the underlying methodology and the description of program AQ11", Machine Learning and Inference Laboratory, George Mason University, 1983 (www.mli.gmu.edu/michalski)
11. Martin Ester, Hans-Peter Kriegel, Jorg Sander, Michael Wimmer, Xiaowei Xu, "Incremental Clustering for Mining in a Data Warehousing Environment" Proceedings of the 24th VLDB conference New York, 1998.