**MODULE – IV**

**DISTRIBUTED DATA MINING**

**Distributed Data Mining**

Data mining technology is used as a mode of identifying patterns and trends from large quantities of data. The Data Mining technology use data integration method to generate Data warehouse, where all data put together on one of the site which is treated as a central site, and then data mining algorithm executed against that data to extract the useful module prediction and knowledge evaluation.

Distributed Data Mining (DDM) is a field which deals with analyzing distributed data and proposes algorithmic solutions to perform different data analysis and mining operations in a distributed manner by considering the resource constraints.

The developments in data mining leads towards distributed data mining (DDM) that mines data sources, regardless of their physical locations, which consists of several and different sources of large volumes of data and several computing units. The most common and prominent example of a distributed environment is the Internet, where increasingly more databases and data streams appear that deal with several diversified areas. Also the Internet use as communication media for geographically distributed information systems. Other examples of distributed mining are process monitoring using sensor networks and grids for the system where a large number of computing and storage units are interconnected over a high speed network.

In short Distributed Data Mining (DDM) is to extract useful information, knowledge and patterns from distributed heterogeneous data bases i.e., to compose them within a distributed knowledge base and use for the purposes of decision making.

**Approaches of data mining algorithms**

The existing data mining algorithms can work in three different computing environments:

* Centralized
* Parallel
* Distributed

**Centralized approach for data mining**

* In Centralized approach, data is extracted and accumulated on a centralized store after cleaning and pre processing. From this central store, task relevant data is selected and mining techniques are applied.
* The centralized algorithm are C4.5, K-means, SVM, Apriori, EM, Page rank

**Parallel approach for data mining**

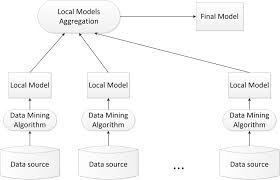
* Many scientific and large problems can be better solved using parallel programming approach
* Data mining can be executed in a high parallel environment over multiple processors.
* Parallel implementations of data mining algorithm can be distinguished on the basis of task parallel and data parallel approaches.

**Distributed approach for data mining**

* Distributed approach for data mining is useful when the data sources are at multiple sites.
* Data extraction, cleaning, pre processing and integrating consumes majority of time there by affecting the analysis process. When it comes to time critical applications, this delay cannot be tolerated. Thus, there exists a requirement to mine such data in a distributed manner.

**Distributed Data Mining Framework**

The following figure shows the Distributed Data Mining framework. In this the data sources are distributed physically and each data source can be mined separately with a data mining algorithm and a local model can be developed. The outputs of all the local models will be aggregated and a final model will be generated.

****

**Distributed Data Mining Framework**

**Distributed Data Source**

In a distributed setting, the data are distributed across several data sources. Each data source contains only a fragment of the data. This leads to the fragmentation of data. Two common types of data fragmentation are horizontal fragmentation - wherein (possibly overlapping) subsets of data tuples are stored at different sites and vertical fragmentation - wherein (possibly overlapping) subtuples of data tuples are stored at different sites. More generally, the data may be fragmented into a set of relations (tables of a relational database, distributed across multiple sites).

Distributed data sources are divided into two types

* Homogeneous - Contains the same set of attributes across distributed data sites
* Heterogeneous - Contains different sets of attributes across distributed data sites.

**Distributed Data Mining techniques**

Distributed data mining (DDM) techniques are necessary for large and multi-scenario datasets requiring resources, which are heterogeneous and distributed. The following are the distributed data mining techniques:

* Distributed classifier learning
* Distributed clustering
* Distributed association rule mining

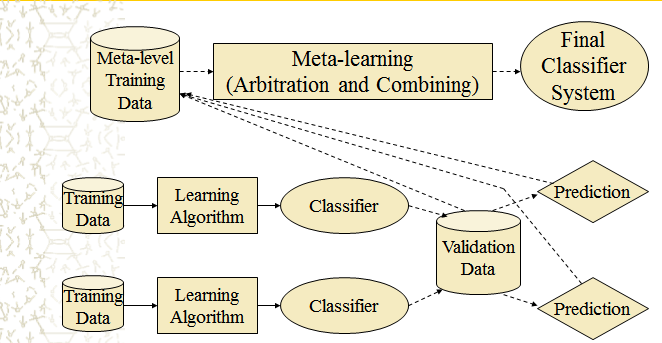
1. **Distributed Classifier learning**

Distributed classifier learning is the technique used to perform classification of data in the distributed environment. Classification is the process of dividing the data into known groups. There are two approaches in distributed classifier learning. They are:

* Meta - learning framework
* Distributed learning with knowledge probing

**Meta – learning framework**

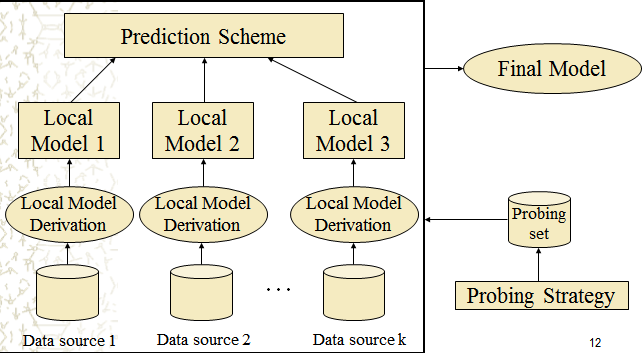
Meta-learning classifier approach denotes use of meta-classifier and base-classifier. This classifier approach proved to be effective, scalable, portable, compatible, extensible and efficient. Meta-learning includes both arbitration and combining. Arbitration generates final prediction result of the feature vector. Combining generates final prediction based on classifier output and classification output or based on classifier output, classification output and feature-vector prediction.



**Meta – learning framework**

**Distributed learning with knowledge probing**

Knowledge-probing is defined as combining several local models to generate the final global model. Steps involved in knowledge-probing include generating base-classifier from an off-the-shelf classifier model, selecting untagged data for probe set, preparing probe set by accumulating final result from base-classifier and finally generating final prediction model of the probe data set. The main difference between knowledge-probing and meta-learning is: knowledge-probing is relying on probe data set for its final prediction, whereas meta-learning involves arbitration and combining learning methods for the final prediction. Knowledge probing in DDM is a two stage process. A collection of base-classifiers is trained in first stage. In second-stage meta-learning technique is applied to attributes for prediction. Further, collaboration of knowledge-probing and other traditional DM algorithms resulted in improved performance. This approach is applied for homogeneous data sources.

****

**Distributed learning with knowledge probing**

1. **Distributed clustering**

Clustering is an unsupervised learning method group objects into clusters such that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters.

The main objective of distributed clustering algorithms is to cluster the distributed datasets without necessarily downloading all the data to the single site. It assumes that the objects to be clustered reside on different sites. This process is carried out on two different levels: local level and global level. On the local level, all sites carry out clustering process independently from each other. After having completed the clustering, a local model or local representative is determined, which should reflect an optimum trade-off between complexity and accuracy. Next, the local model is transferred to a central site, where the local models are merged in order to form a global model. The global model is again transmitted to local sites to update local models. Distributed clustering algorithms are mostly applied on homogeneous data sites.

Eg: Distributed k-means algorithm

1. **Distributed association rule mining**

The distributed association rule mining is used to discover all rules with global support and global confidence greater than the user specified minimum support and confidence.

Let us consider; in a distributed data mining environment collective database DB is subdivided into DB1, DB2, … , DBN in collection of data sites S1, S2, … ,SN respectively. I= {i1, i2, … , im} is the set of items where each transaction T⊆I. Typical form of an association rule is X⇒Y, where X⊆I, Y⊆I and X∩Y=φ. The support s of X⇒Y is the probability of a transaction in DB containing both X and Y. On the other hand confidence c of X⇒Y is the probability of a transaction containing X will contain Y too. Usually it is the interest of the data miners to find all association rules having support and confidence greater than or equal to minimum threshold value.

The following are some of the distributed association rule mining algorithms:

**Count Distribution (CD)**

Count Distribution (CD) algorithm works as follows: Each processor or data site generates its local candidate sets based on the global large itmesets of previous iteration using Apriori algorithm. Then it calculates each support count and exchanges with other sites using Message Passing Interface (MPI) technique. Since this protocol exchanges all counts, each site can generate global frequent large itemsets which might be utilized for the following iterations. Due to the use of the same algorithm, all the processors generate same global frequent large itemsets. CD algorithm can be summarized into five major stages:

1. Each processor generates candidate itemset Ck based on globally frequent large itemset Lk-1.
2. Each processor computes local support count for Ck by passing through the transactions in the database.
3. All processors exchange their Ck counts to develop global Ck using MPI technique.
4. Each processor computes Lk from Ck.
5. Each processor takes the decision either to continue or to stop. Decision will be the same since they have identical Lk.

**Fast Distributed Algorithm (FDM)**

Fast Distributed Mining of Association Rules (FDM), was proposed by W. Cheung, J. Han, V. T. Ng,A. W. Fu, Y. Fu (1996). The main idea of this protocol can be summarized as follows:

1. Computing candidate set: Each site generates candidate set based on globally large (k-1)-itemsets and locally large (k-1)-itemsets using Apriori algorithm.
2. Local pruning: For each item in the candidate set: if the support of the itemset is larger than minimum support, that particular item is added in the locally large k-itemsets.
3. Count exchange: Each site broadcasts locally frequent large itemsets to all other sites.
4. Globally frequent large itemset computation: Each site computes globally large k-itemsets which is utilized for the following iteration.

**Challenges of Distributed Data Mining**

The major issues that affect the performance of Distributed Data Mining are as follows:

**Heterogeneous data mining:** If the data is heterogeneous then contradiction among the attributes will occur. Also if the data is heterogeneous; local data management model should be integrated into a global model before dealing with the data items.

**Data consistency:** Since data is distributed across many sites, it creates a problem of data inconsistency. The modification applied on in local data model if not reflected to global data model or global database it may affects the final result produced after Data Mining.

**Communication cost:** Communication cost depends on network bandwidth and amount of information transferred. In Distributed Data Mining a cost model should be built.

**Knowledge integration:** Knowledge Integration deals with integrating local results to produce global results. It is the critical step in any Distributed Data Mining. During the integration process the local models should not lose its value in the global range, it must be preserved.

**Data variance:** In distributed environment data is not static as that of traditional data mining. Along with data the executing environment is also dynamic; hence the Distributed Data Mining algorithm should correctly transfer the time series result and time series related result.

**Privacy preserving:** The main objective of privacy preserving data mining is to develop algorithm for modifying original data in some way so that private data and private knowledge remain private even after mining process. The problem that occurs when unauthorized user derives confidential information from released data which is commonly called database inference problem. Most recent efforts towards addressing the privacy issue are data distortion and cryptographic methods.