I J C T A, 9(3), 2016, pp. 1517-1527

Clustered Collaborative Filtering Approach for Distributed Data Mining on Electronic Health Records

Dr. M. Nandhini*, S. Urmela**

Abstract: Distributed Data Mining (DDM) has become one of the promising areas of Data Mining. DDM techniques include classifier approach and agent-approach. Classifier approach plays a vital role in mining distributed data, having homogeneous and heterogeneous approaches depend on data sites. Homogeneous classifier approach involves ensemble learning, distributed association rule mining, meta-learning and knowledge probing. Heterogeneous classifier approach involves collective principle component analysis, distributed clustering, collective decision tree and collective bayesian learning model. In this paper, classifier approach for DDM is summarized and an architectural model based on clustered-collaborative filtering for Electronic Health Records (EHR) is proposed.

Keywords: Distributed Data Mining, classifier approach, clustering, collaborative filtering, Electronic Health Record, clustered collaborative filtering.

1. INTRODUCTION

Data mining is the process of extracting useful, unknown information, from data in databases using patterns. The progressive growth of information and technology has paved way to further explore Distributed/ Collective Data Mining, Spatial and Geographic data mining, Temporal data mining, Spatio-Temporal data mining, Multimedia data mining and phenomenal data mining. Data mining today performs computation on database or warehouse at a single geographical location. Future scope of data mining involves computing data located at different geographical locations. This is termed DDM/CDM. The objective of DDM is to extract useful, unknown information from data located at heterogeneous sites. Distributed computing involves distributed sites, hosting computing units at each heterogeneous points[1].

The main factors which led to evolution of DDM are – privacy of sensitive data, transmission cost, computation cost and memory cost. DDM follows decentralized mining strategy which differs from centralized strategy making entire working system scalable by distributing workload across heterogeneous sites. Further, following centralized strategy involves data prone to security and privacy risks[1].

Decentralized/Distributed strategy involves data storage at heterogeneous sites, thereby lessening security attacks and providing Confidentiality, Integrity and Availability of useful information. DDM mainly involves two variations—data distributed and computation distributed. In former method, data will be distributed among heterogeneous sites at local level and computation will be hosted at global level. In latter method, computation will be distributed among heterogeneous sites at local level and data will be hosted at global level. [1].

Figure 1.1 explains DDM working architecture. The database at heterogeneous sites hosts useful, unknown information. DDM algorithms will be applied over data at heterogeneous sites as local model and finally the data mining computed result will be agglomerated to form global model[1].

* Assistant ProfessorDepartment of Computer Science Pondicherry University, *Email: mnandhini2005@yahoo.com*

** Ph.D Research ScholarDepartment of Computer Science Pondicherry University, *Email: urmelaindra@gmail.com*

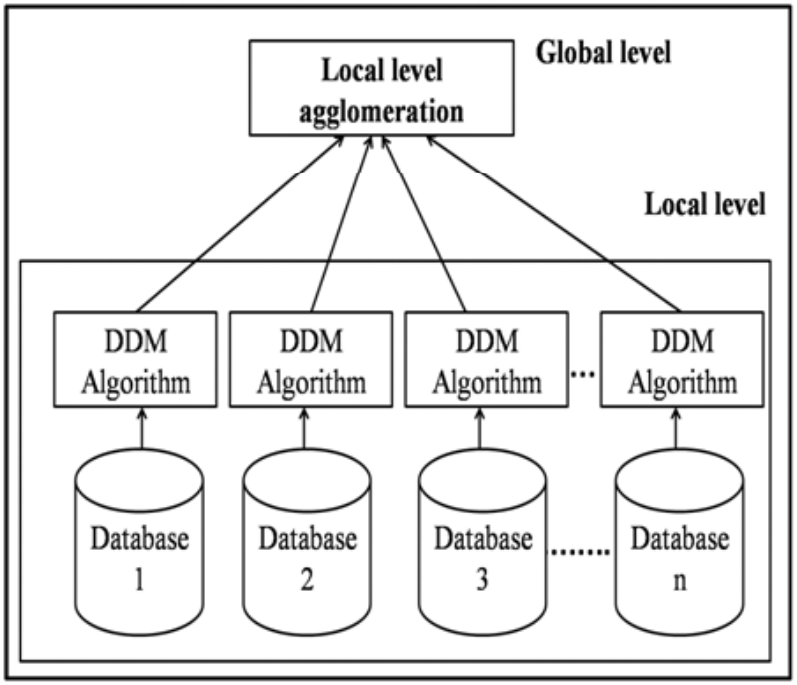

Figure 1: Working Architecture – Distributed Data Mining

Yan Li et al[2] proposed a novel privacy-based distributed ensemble classifier approach for predicting model for EHR data. Each participating homogeneous sites will accumulate dataset in local level. Finally, at global level prediction model will be generated from multiple local models. Iyad Batal et al[3] proposed a framework based on temporal pattern. The framework is able to make decision-making by retrieving knowledge by data mining. The proposed work involves decision-making and patients' record management tasks.

Mining on EHR in centralized environment paves way for increased medical cost interms of repeated laboratory tests and degrades promotion of effective clinical decision-making. This paper involves proposal of mining EHR in distributed environment.

The organization of the paper is as follows: Section II describes an overview of DDM based on classifier approach. Section III discusses the related work on DDM based on classifier approach and DDM on EHR. Section IV depicts an abstract model for DDM on EHR. Section V summarizes the paper.

2. DISTRIBUTED DATA MINING BASED ON CLASSIFIER APPROACH

Distributed Data Source

Based on distributed data source, DDM can be classified into two approaches:[4]

i) Homogeneous Classifier approach

ii) Heterogeneous Classifier approach

Homogeneous Classifier approach

In this classifier approach, the database will be maintaining same set of attributes across distributed geographical sites.

i) Heterogeneous Classifier approach

In this classifier approach, the database will be maintaining different set of attributes across distributed geographical sites.

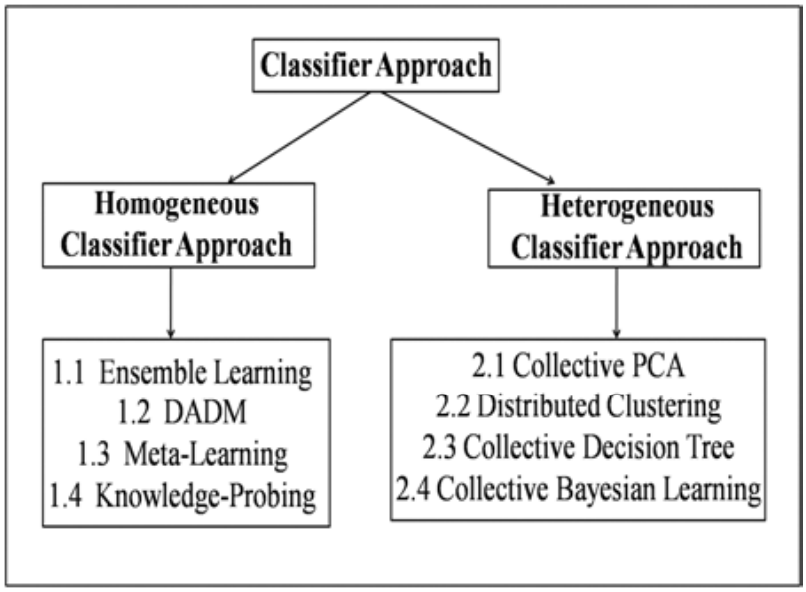

Figure 2: Classifier approach-Distributed Data Mining[4]

Now let's take a look at each classifier approach in detail. Some approaches will be similar to data mining algorithms.

1. Homogeneous Classifier approach

Homogeneous classifier approach, maintain same set of attributes across distributed geographical sites.

1.1. Ensemble Learning

An Ensemble Learning approach involves multiple learning models to obtain final predictions. An ensemble learning classifier approach proves to be an effective learning approach, in-terms of combining multiple learning models giving better prediction result than any of the solo classifier approach. Some of the well-known Ensemble Learning classifier approaches involve bagging, boosting, random forest, stacking and arcing[5].

Out of these five ensemble learning classifier approach, bagging and boosting proves to be effective ensemble learning classifier approach[5].

1.2. Distributed Association Rule Mining (DARM)

DARM involves certain association rules for generating local datasets. Finally the global datasets will be generated from multiple local datasets. There are three algorithms involved in DARM[5],

i) Count Distribution algorithm involves Apriori algorithm generating k-itemsets for each iteration at local level, global level computes the final-itemsets[5][6].

ii) Fast DARM algorithm involves pruning of itemsets at local level where pruning is followed for each iteration [5][6].

iii) Optimized DARM algorithm involves both Count Distribution algorithm and Fast DARM algorithm. It performs efficiently than former two algorithms by deleting earlier itemsets at local level and deleting duplicate transactions by keeping track of a counter[5][6].

1.3. Meta-Learning

Meta-Learning classifier approach involves use of meta-classifier and base-classifier. This classifier approach proves to be effective, scalable, portable, compatible, extensible and efficient[5].

Meta-Learning involves arbitration and combining. Arbitration involves final prediction result from the feature vector. Combining involves final prediction based on classifier output and classification output or based on classifier output, classification output and feature-vector prediction[5].

1.4. Knowledge-Probing

Knowledge-Probing involves combining several local models to generate final global model. Steps involved in Knowledge-Probing include generating base-classifier from 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 from the probe data set[5].

The main difference between Knowledge-Probing and Meta-Learning is: Knowledge-Probing will be relying on probe data set for its final prediction, whereas Meta-learning involves arbitration and combining learning methods for final prediction[5].

2. HETEROGENEOUS CLASSIFIER APPROACH

As previously discussed Heterogeneous classifier approach, will be maintaining different set of attributes across distributed geographical sites.

2.1. Collective Principle Component Analysis[4]

Collective PCA involves Heterogeneous classifier approach, by performing PCA on local dataset, by selected eigen vector set. Finally global dataset prediction involves, combining selected dominant eigen vector set obtained by PCA on dataset[5].

2.2. Distributed Clustering[4]

Distributed Clustering, Heterogeneous classifier approach involves three approach,

i) Collective Hierarchical Clustering (CHC) algorithm[6],

ii) Recursive Agglomeration of Clustering Hierarchies by Encircling Tactic **(**RACHET) algorithm[6],

iii) Density Based Distributed Clustering (DBDC) algorithm[5][6].

i) *Collective Hierarchical Clustering (CHC)*

This algorithm involves dendrogram, a tree representation of clusters. Local dendrograms will be generated at each local geographical site. Finally global dendrogram will be generated from multiple transmitted local dendrograms[5][6].

ii) *Recursive Agglomeration of Clustering Hierarchies by Encircling Tactic (RACHET)*

Hierarchical clustering algorithm will be generated at each local geographical site; separate statistics set will be generated for each site. Finally global model agglomerates local dendrogram to generate final predictions[5][6].

iii) *Density Based Distributed Clustering (DBDC)*

DBSCAN algorithm will generate local cluster prediction model, at each heterogeneous local site. Representative points of each cluster sets will be selected and finally they will combine at global level for final prediction[5][6].

2.3. Collective Decision Tree[4]

Collective Decision Tree, Heterogeneous classifier approach involves Decision Tree model generation at local geographical heterogeneous site. Finally, global level prediction involves collection of local Decision Tree models[5].

2.4. Collective Bayesian Learning[4]

Collective Bayesian Learning, Heterogeneous classifier approach involves Bayesian learning model generation at local geographical heterogeneous site. Finally, global level prediction involves collection of local Bayesian learning models[5].

3. RELATED WORKS

Some related works in the field of DDM by classifier approach is discussed here.

A. By Yan Li et al, "**A distributed ensemble approach for mining health care data under privacy constraints**[2]", involves proposal of a novel privacy-based distributed ensemble classifier approach, for predicting model for EHR data. Each participating homogeneous sites will accumulate dataset in local level. Finally, at global level prediction model will be generated from multiple local models. Main advantage of this proposal is less computational complexity and communication cost.

B. By Hemanta Kumar Bhuyan et al, **"Privacy preserving sub-feature selection in distributed data mining**[7]", involves sub-feature selection by fuzzy method, thereby maintain privacy of original data. Two-fuzzy sets are generated using borelset, helps in determining sub-feature selection within certain interval. The work shows effective and better performance compared to traditional methods. Privacy of original data is maintained. Main advantage of this proposal is efficient sub-feature selection and privacy of original data.

C. By Kawuu W. Lin et al, "**A fast and resource efficient mining algorithm for discovering frequent patterns in distributed computing environments**[8]", involves automatic allocation of local-level nodes for detecting frequent patterns. Previous methods involve initially assigning computing nodes for each transaction thereby, decreasing load-balancing effect. Proposed, mining algorithm doesn't involve any parameter but still able to discover patterns, without initially setting required number of nodes. Main advantage of this proposal is efficient load-balancing, execution efficient and network transmission cost.

D. By A.O. Ogunde et al, "**A partition enhanced mining algorithm for distributed association rule mining systems**[9][10][11]", involves association rule mining agent assigning coordinating agents, which receives request and determines the required geographical sites.

E. By Dr. C.Sunil Kumar et al, "**An Apriori Algorithm in Distributed Data Mining System**[12]", involves distributed mining on XML data. Since mining XML data is difficult, the proposed algorithm, ODAM (Optimal Association Rule Mining), involves mining process in parallel. It achieves better response time and minimized communication cost.

F. By Trilok Nath Pandey et al, "**Improving performance of distributed data mining (DDM) with multi-agent system**[13][14][15][16][17][18]", involves improving DDM performance by Mobile-agent which involves query optimization, discovery plan, local knowledge discovery and knowledge consolidation. Main advantage of this proposal is accurate information retrieval and decreased communication and memory overhead. Privacy of original data is compromised.

G. By Kamalika Das et al, "**A local asynchronous distributed privacy preserving feature selection algorithm for large peer-to-peer networks**[19]", involves feature selection in asynchronous

manner, having decreased communication overhead, thereby maintaining privacy of original data. Each participating node collects data from local level nodes. At global level, final model will be generated from multiple local models. Main advantage of this proposal is scalability, accurate, privacy of original data. Computational complexity is increased.

H. By Frank S.C. Tseng et al, "**Toward boosting distributed association rule mining by data de-clustering**[20]", involves data de-clustering, by which datasets will be de-clustered into partitions. Round-robin method will be followed for iterative assigning of dataset to participating geographic data sites. Load-balancing approach is followed where itemsets of each geographic site will be generated quickly. Main advantage of this proposal is decreased communication cost and space complexity.

I. By Golam Kaosar et al, "**Distributed Association Rule Mining with Minimum Communication Overhead**[21]", involves message passing interface and generating global frequent large itemsets. This proposal involves association rule mining. Pruning techniques helps to reduce communication overhead across distributed geographical data sites. Main advantage of this proposal is decreased communication overhead. Efficiency and privacy of original data is compromised.

J. By Philip K. Chan et al, "**Distributed Data Mining in Credit Card Fraud Detection**[22]", involves detecting fraud credit card transactions by maintaining frequent patterns of transactions across distributed geographic sites. The proposed method involves scalable and efficient approach, by generating learning model in base-classifiers. Meta-learning classifier approach is followed; base-classifier involves predictive learning models obtained from meta-classifier.

Since learning models are used for prediction several base-classifiers at each geographical site can operate in parallel with meta-classifier. Highly-skewed data has been studied in this approach. Main advantage of this proposal is scalability, efficient and cost-effective solution. Implementation of adaptive approach is the main disadvantage of this approach.

Table 3.1 depicts comparison work of DDM based on classifier approaches along with the methodology followed and its pros and cons.

Table 3
Classifier Approach-Distributed Data Mining

Title	*Author*	*Classifier approach*	*Methodology*	*Pros*	*Cons*
A distributed ensemble approach for mining health care data under privacy constraints [2]	Yan Li et al.(2016)	**Ensemble Learning (Boosting) Homogeneous approach (Heterogeneous data bridging)**	• Adaptive distributed privacy-preserving data mining by AdaBoost	• Communication cost lesser compared to star network (untrusted third party) • Complexity is lesser when new participator is added	• Memory overhead in learning other participator models
Privacy preserving sub-feature select-ion in distributed data mining[7]	Hemanta Kumar Bhuyan et al. (2015)	**Heterogeneous approach (Fuzzy model)**	• Sub-feature selection involves fuzzy methodology • Borel set generates two fuzzy set which determines sub-feature selection	• Efficient sub-feature selection approach • Privacy of original data	• During develop-ing fuzzy membership function outlier values are discarded

(*contd...*)

(*Table 3 contd...*)

Title	*Author*	*Classifier approach*	*Methodology*	*Pros*	*Cons*
A fast and resource efficient mining algorithm for discovering Frequent patterns in distributed computing environments[8]	Kawuu W. Lin et al. (2015)	**Distributed Association mining Homogeneous approach**	• FLR-algorithm iteratively determines number of computing nodes for mining process	• Load-balancing • Execution efficiency • Parameter-less, less manual interaction • Network transmission cost minimized	• Privacy of original data
A partition enhanced mining algorithm fordistributed association rule mining systems[9][10][11]	A.O. Ogunde et al.(2015)	**Distributed Association mining Homogeneous approach (Agent-based)**	• Partition Enhanced Mining Algorithm, involves logic of mobile-agent based vertical partitioning very large data into distributed data sites	• Reduced response time • Communication cost • Scalability • efficiency	• Use in heterogeneous environments • Security issues

Title	*Author*	*Classifier approach*	*Methodology*	*Pros*	*Cons*
An Apriori Algorithm in Distributed Data Mining System[12]	Dr. C. Sunil Kumar et al.(2013)	**Distributed Association mining Homogeneous approach**	• Association Rule mining on XML data by load-balancing	• Reduced response time • Execution efficiency	• Privacy of original data
Improving performance of distributed data mining (DDM) with multi-agent system[13][14][15][16][17][18]	Trilok Nath Pandey et al.(2012)	**Distributed Clustering Heterogeneous approach**	• Mobile-agent based Distributed mining involves query optimization, discovery plan, local knowledge discovery and knowledge consolidation	• Accurate • Communication overhead	• Most expensive • Privacy of original data
A local asynchronous distributed privacy preservingfeature selection algorithm for large peer-to-peer networks[19]	Kamalika Das et al.(2010)	**Bayesian model Heterogeneous approach**	• Two algorithms (L-ring and PAFS) • Feature selection in an asynchronous manner • Each peer decides its own privacy constraints, local interaction among participating nodes	• Scalable • Accurate • Communication overhead • Privacy preserving of original data	• Computational complexity
Toward boosting distributed association rule mining by data de-clustering[20]	Frank S.C. Tseng et al.(2010)	**Distributed Association rule mining Homogeneous approach**	• Shortest spanning path used to decluster dataset into subgroups among participating nodes	• Participating nodes communication cost • Space complexity	• Distributed database declustering
Distributed Association Rule Mining with Minimum Communication Overhead[21]	Golam Kaosar et al.(2009)	**Distributed Association mining Homogeneous approach**	• Fast Distribution algorithm involves message passing interface and generate global	• Communication overhead	• Efficiency • Privacy of original data

(*contd...*)

(*Table 3 contd...*)

Title	*Author*	*Classifier approach*	*Methodology*	*Pros*	*Cons*
			frequent large itemsets		
Distributed Data Mining in Credit Card Fraud Detection[22]	Philip K. Chan et al.(1999)	**Meta Classifier (Learning) Homogeneous approach (Heterogeneous data bridging)**	• Grouping Datasets (legitimate or fraud) • At Distributed site mining technique applied over base-classifier • Base models combined to formulate meta classifier	• Scalable • Efficient • Highly skewed data considered • Cost-based mining tech	• Implementation of (adaptive-Time-dependent) Classifier approach

Though most of the existing works focuses mainly on homogeneous classifier approach privacy of original data, computational complexity and memory overhead are certain cons. Main aim of DDM is to maximize privacy of original data, minimize computational complexity and memory overhead. Our proposed approach on EHR in this paper concentrates on minimizing computational complexity and memory overhead.

Some related works in the field of DDM on EHR is discussed here.

A. By Yan Li et al, "**A distributed ensemble approach for mining health care data under privacy constraints** [2]", involves proposal of a novel privacy-based distributed ensemble classifier approach, for predicting model for EHR data. Each participating homogeneous sites will accumulate dataset in local level. Finally, at global level prediction model will be generated from multiple local models. Main advantage of this proposal is less computational complexity and communication cost.

B. By Shaker H. El-Sappagh et al, "**Electronic Health Record Data Model Optimized for Knowledge Discovery**[23]", involves proposal of an abstract data model by relational object data model. The model uses class and relationship attributes. The proposed work involves decision-making and mining patients' records. Main advantage of this proposal is problem-oriented EHR. Interoperability of patients' record is an issue to be further explored.

C. By Iyad Batal et al, "**A Temporal Pattern Mining Approach for Classifying Electronic Health Record Data**[3]", involves proposal of framework based on temporal pattern. The framework is able to make decision-making by retrieving knowledge by data mining. The proposed work involves decision-making and patients' record management tasks.

D. By David Gotz et al," **A methodology for interactive mining and visual analysis of clinical event patterns using electronic health record data**[24]", involves visual query pattern mining of EHR. The model involves an interactive visual query pattern mining by event pattern analysis.

Above discussed works of DDM on EHR, mainly focuses on decision support and record management tasks. The goal of DDM on EHR is to minimize computational complexity and memory overhead. The proposed architecture of clustered collaborative filtering for DDM on EHR mainly focuses on minimizing computational complexity and memory overhead apart from promoting effective clinical decision-making and efficient EHR management.

4. ARCHITECTURE–CLUSTERED-COLLABORATIVE FILTERING FOR DISTRIBUTED DATA MINING ON EHR

In this section architecture as in Figure 3 is proposed with clustered-CF for DDM on EHR. The entities and desires of EHR and the procedure for clustered-CF for DDM is explained below. The proposed clustered-CF for DDM on EHR, working collaboratively will result in less memory overhead and decreased computational complexity.

A. Electronic Health Record

EHR's are patient-oriented records, makes information readily available to legitimate users. EHR includes patients' therapeutic history, immunization date, drugs, allergies and test outcomes. Benefits of EHR include manual error avoidance, timely notification of patient information like immunization date, appointments, minimizing allergies to certain drug effects[25].

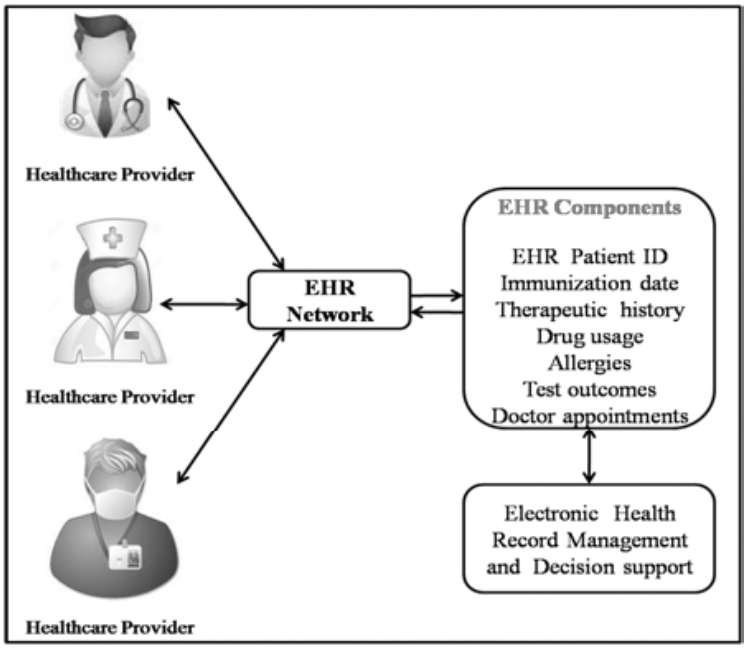

Figure 3: EHR-Work Flow

B. CLUSTERED-COLLABORATIVE FILTERING

CF involves decision based on previous record history. In this case CF approach will be applied to diagnose a patient based on previous patients' record with same symptoms or other purpose. Each distributed data site, considering as cluster will involve CF. CF algorithm involves memory-based[26] and model-based[27] approach. Memory-based involves user-user similarity technique which accurately identifies patients' record. Model-based involves bayesian network, clustering and rule association technique which initially models dataset based on Bayesian network model, clustering items and association respectively. Further a third variation of CF (hybrid-CF) is discussed which involves both memory-based CF and model-based CF. The way in which both CF has combined is managed by hybrid-switching technique[28].

In the proposed work, we have gone for meta-level hybrid switching technique[28] on which a model formulated for model-based CF will be applied over memory-based CF. By which memory-based CF involves forming cluster of distinct disease identified from history of EHRs (old patients EHRs). Model-based CF on new patients' EHRs retrieval was done. EHRs retrieval by keyword based will be searching for EHRs in

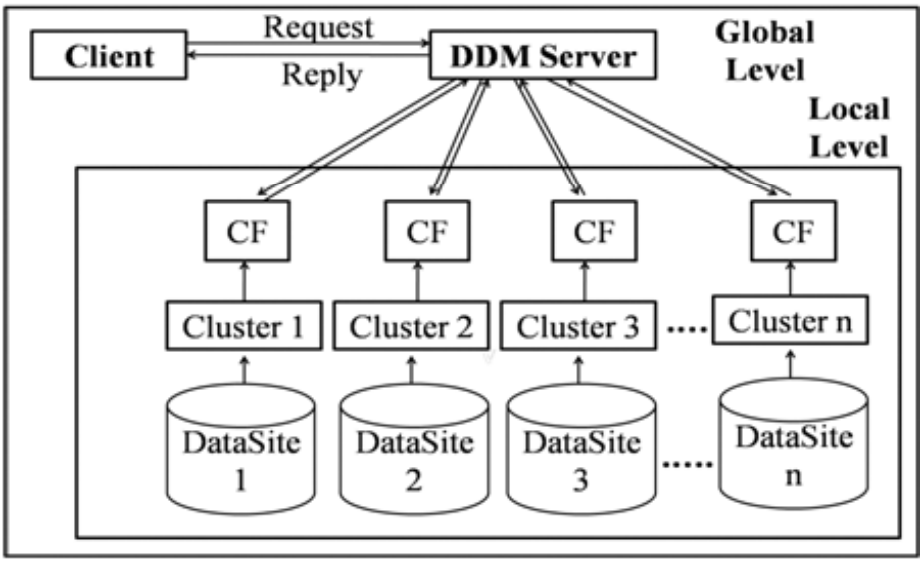

Figure 4: Architecture-Clustered-Collaborative Filtering for Distributed Data Mining on EHR

entire dataset sequentially leading to maximum computational complexity whereas the proposed EHRs retrieval architecture by memory-based CF and model-based CF for DDM has minimized computational complexity and memory overhead by forming clusters of EHRs from history of EHRs.

C. CCF ON EHR

The proposed architecture of Clustered-CF on EHR involves CF of patients' historic records for diagnosing.

At local level of DDM, based on client request to DDM server, each distributed site will be collaboratively filtering out patient information by user-user similarity of patients' record. The retrieved patients' record will be accumulated using K-Means Clustering algorithm. At global level of DDM, the retrieved data from DDM server will be sent to client as a response.

This proposed clustered-CF for DDM on EHR is expected to have less memory overhead and decreased computational complexity.

5. CONCLUSION

A brief summary and survey on DDM-classifier approach for different applications are given over a period of years. Further, an abstract model for clustered-CF is proposed for DDM on EHR. This enables us to diagnose patients' medical record efficiently and accurately by minimizing computational complexity and memory overhead.

References

[1] Ms. Vinaya Sawant, Dr. Ketan Shah, "A review of Distributed Data Mining using agents", International Journal of Advanced Technology & Engineering Research (IJATER), Volume 3, issue 5, September 2013, pp. 27-33.

[2] Yan Li, Changxin Bai, Chandan K. Reddy, "A distributed ensemble approach for mining health care data under privacy constraints", Journal of Information Sciences, Volume 330, February 2016, pp. 245-259.

[3] Iyad Batal, Hamed Valizadegan, Gregory F. Cooper, Milos Hauskrecht, "A Temporal Pattern Mining Approach for Classifying Electronic Health Record Data", ACM Transactions on Intelligent Systems and Technology, Volume 4, Issue 4, August 2012.

[4] Hillol Kargupta, "An Introduction to Distributed Data Mining", http://www.eecs.wsu.edu/~hillol

[5] S. V. S. Ganga Devi, "A Survey On Distributed Data Mining And Its Trends", International Journal of Research in Engineering & Technology (IJRET), Volume 2, Issue 3, March 2014, pp. 107-120.

[6] G *Tsoumakas*, E Spyromitros-Xioufis, J Vilcek, I Vlahavas "Distributed Data Mining", Proc.ECML/PKDD, Workshopon *Mining* Multidimensional *Data* (MMD'08), 30-44, 2008.

[7] Hemanta Kumar Bhuyan, Narendra Kumar Kamila, "Privacy preserving sub-feature selection in distributed data mining", Journal of Applied Soft Computing Volume 36, November 2015, pp. 552-569.

[8] Kawuu W. Lin, Sheng-Hao Chung, "A fast and resource efficient mining algorithm for discovering Frequent patterns in distributed computing environments", Journal of Future Generation Computer Systems, Volume 52, November 2015, pp. 49-58.

[9] A.O. Ogunde, O. Folorunso, A.S. Sodiya, "A partition enhanced mining algorithm for distributed association rule mining systems" Egyptian Informatics Journal, Volume 16, Issue 3, November 2015, pp. 297-307.

[10] http://www.sciencedirect.com/science/article/pii/S1110866515000365

[11] https://www.researchgate.net/publication/260386044_DARCI_Distributed_Association_Rule_Mining_Utilizing_Closed_Itemsets

[12] Dr. C.Sunil Kumar, P.N.Santosh Kumar & Dr. C.Venugopal, "An Apriori Algorithm in Distributed Data Mining System", Global Journal of Computer Science and Technology Software & Data Engineering, Volume 13, Issue 12, 2013.

[13] Trilok Nath Pandey, Niranjan Panda, Pravat Kumar Sahu, "Improving performance of distributed data mining (DDM) with multi-agent system", (IJCSI) International Journal of Computer Science Issues, Vol. 9, Issue 2, No 3, March 2012, pp. 74-82.

[14] http://connection.ebscohost.com/c/articles/75184256/improving-performance-distributed-data-mining-ddm-multi-agent-system

[15] http://www.caeaccess.org/archives/volume3/number8/486-2015652008

[16] http://www.techrepublic.com/resource-library/whitepapers/improving-performance-of-distributed-data-mining-ddm-with-multi-agent-system/

[17] http://www.jourlib.org/paper/2340249#.Vp-VwZp97IU

[18] http://www.engpaper.com/data-mining-research-papers-2012-section-7.htm

[19] Kamalika Das, Kanishka Bhaduri, Hillol Kargupta, "A local asynchronous distributed privacy preserving feature selection algorithm for large peer-to-peer networks", Journal of Knowledge and Information Systems, Volume 24, Issue 3, September 2010, pp. 341-367.

[20] Frank S.C. Tseng, Yen-Hung Kuo, Yueh-Min Huang, "Toward boosting distributed association rule mining by data de-clustering", Journal of Information Sciences, Volume 180, Issue 22, November 2010, pp. 4263-4289.

[21] Md. Golam Kaosar, Zhuojia Xu, Xun Yi, "Distributed Association Rule Mining with Minimum Communication Overhead", Proc. of the 8th Australasian Data Mining Conference (AusDM'09), Volume 101, pp. 17-23.

[22] Philip K. Chan, Wei Fan, Andreas L. Prodromidis, Salvatore J. Stolfo, "Distributed Data Mining in Credit Card Fraud Detection", IEEE Intelligent Systems, December 1999, pp. 67-74.

[23] Shaker H. El-Sappagh, Samir El-Masri, A. M. Riad, Mohammed Elmogy, "Electronic Health Record Data Model Optimized for Knowledge Discovery", IJCSI International Journal of Computer Science Issues, Volume 9, Issue 5, No 1, September 2012, pp. 329-338.

[24] David Gotz, Fei Wang, Adam Perer, "A methodology for interactive mining and visual analysis of clinical event patterns using electronic health record data", Journal of Biomedical Informatics Volume 48, April 2014, pp. 148–159.

[25] https://www.healthit.gov/providers-professionals/faqs/what-electronic-health-record-ehr

[26] Xiaoyuan Su, Taghi M. Khoshgoftaar, "A Survey of Collaborative Filtering Techniques", Hindawi Publishing Corporation, Advances in Artificial Intelligence, Article ID 421425, 19 pages, http://dx.doi.org/10.1155/2009/421425.

[27] Y. Koren, "Tutorial on recent progress in collaborative filtering", Proceedings of the the 2nd ACM Conference on Recommender Systems, 2008.

[28] A survey of Collaborative Filtering Techniques: http://www.hindawi.com/journals/aai/2009/421425/