ARTICLE

CBPLSA - AN EFFECTIVE COLLABORATIVE FILTERING ALGORITHM FOR DISTRIBUTED DATA MINING ON ELECTRONIC HEALTH RECORDS

S. Urmela[1*], M. Nandhini[2]

[1]*Department of Computer Science, Pondicherry University, Puducherry, INDIA*

ISSN 0976-3104

ABSTRACT

Background: *In recent years, Distributed Data Mining (DDM) on Electronic Health Records (EHRs) has become one of dominant area in DM.* ***Methods:*** *This paper proposed architecture for EHRs DDM using memory-based and model-based collaborative filtering. The proposed CBPLSA(Cluster-Based Probabilistic Latent Semantic Analysis) algorithm for DDM on EHRs which aims to minimize computational complexity and memory overhead by maintaining clusters of old patients' EHRs.* ***Implementation:*** *Experimental implementation on real-world EHRs datasets available on Hypertension, Diabetes and Meningitis depicts an improved precision and result accuracy compared to state-of-arts EHRs retrieval approaches.*

KEY WORDS
Distributed Data Mining
Electronic Health Records
CBPLSA algorithm
Collaborative Filtering
Clustering

INTRODUCTION

Data mining (DM) is the process of extracting useful, important information using patterns from available datasets.[1] discusses that the tremendous growth of information and technology has paved way to explore other forms of data mining which includes Collective Data Mining (CDM/DDM), temporal data mining, phenomenal data mining and visual data mining. DDM is devoted to retrieve patterns from distributed datasets. Centralized data mining performs data computation at a dedicated geographical location. Processing data at centralized site paves way for questioning on privacy of sensitive data, increased computation cost, memory cost and transmission cost. The objective of DDM is to extract useful, unknown information from heterogeneous data sites. It involves computation at heterogeneous points, hosting individual computing units. De-centralized data mining also makes entire system scalable, by distributing the workload among heterogeneous computing points. Traditional algorithms developed for DM mostly devoted for centralized environments has proven to be unsuitable for DDM [1]. Alfredo Cuzzocrea(2013)[2] states that framing a methodology for DDM is challenging not only by distributed environment, but also for its efficient resource sharing and minimizing computational complexity specifications. Kargupta et al. [3] and Zaki M J et al.[4] discussed that several researchers analyzed complexity involved in framing methodology for DDM in two ways:

- Analyzing on effective and efficient usage of computational resources at individual distributed data-sites.
- Performing knowledge discovery at individual distributed data site (local level).
- Aggregating knowledge discovered at global level.

Received: 18 Feb 2018
Accepted: 31 Mar 2018
Published: 8 April 2018

Byung-Hoon Park et al. [5] developed an architecture for DDM where processing takes place locally at individual data sites. Finally data will be accumulated to form global model. Grigorios Tsoumakas et al. [6] presented an architecture for DDM where knowledge acquired from local distributed data sites are accumulated at global level forming a merger site. Fu Y et al.[7] discussed certain issues in developing DDM algorithms namely formulating suitable DDM algorithm for heterogeneity datasets; minimizing computational complexity and space complexity; communication cost; privacy preservation of data at distributed site; data fragmentation; data replication and maintaining local datasets autonomy. Further all these issues are interrelated to each other. This has given pave to many researchers to carry-out their work in this field mainly on EHRs retrieval.

This work is inspiring us to bring out a high-performance contribution towards DDM by framing a suitable retrieval algorithm to more interesting but partially explored areas like health informatics, e-science and bioinformatics. Mainly in recent years, in the field of health informatics, retrieval of EHRs with DDM is targeting about minimizing computational cost and computational complexity. Research works so-far proposed on EHRs retrieval have covered various retrieval approaches by clustering, classification and association [2]. Inspired by this, to minimize the computational and memory costs in retrieval, we propose an algorithm for EHRs retrieval approach with memory-based CF and model-based CF (CBPLSA algorithm) for DDM.

RELATED WORKS

A number of prominent EHR mining in centralized environment has been designed. But the performance of mining in centralized environment alone is too limited and paves way for universal EHR. By implementing universal EHR, increased health care cost in terms of repeated laboratory tests can be avoided, promotes effective clinical-decision making.

***Corresponding Author**
Email:
urmelaindra@gmail.com
Tel.: +91 9791958868

DDM on EHR

Some of the distributed mining approaches on EHR are discussed below.

Mohammed Khalilia et al.[8] discussed a disease prediction framework based on random forest (RF), an ensemble homogeneous classifier approach based on repeated random sampling of trained datasets. National Inpatient Sample (NIS) data obtained through Healthcare Cost and Utilization Project (HCUP) was used for experimentation. Eight chronic diseases were predicted and performance was compared with other ensemble classifier approach, namely bagging and boosting, along with VM (Vector Machine).

Yan Li et al.[9] discussed a novel security-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.

CF on health prediction

Some of the CF approaches on health prediction are discussed below.

Martin Lopez et al.[10] discussed a new CF, Property-based Collaborative Filtering (PBCF), introduced links among semantic properties which handles both users and items. This approach helped to solve CF problems namely, scalability, by enabling developer to build a matrix of how much a feature for one person influences the same feature item for some another person. PBCF was used in health-aware recommender system. It was implemented in HARE (Heath-Aware Recommender)[10], a system introduced to deliver personalized ads. PBCF was implemented in HARE, as a health-aware recommender system.

Davis Darcy A et al.[11] proposed Collaborative Assessment and Recommendation Engine (CARE) which uses patients' medical history and similarity among patients' for predicting future greater disease risks. Also an iterative version, ICARE was devised which uses ensemble, homogeneous classifier approaches for achieving better performance.

Kai Zheng et al.[12] discussed Electronic Medical Record Search Engine (EMERSE), implemented along with collaborative search by preserving knowledge of collected EHR search and circulated among other EHR searches. This approach involves use of social-gathered information, helping in accurate and efficient health care information retrieval. Complex search terms and smaller EHR search engine size are certain cons. Above discussed works of DDM on EHRs, focus mainly on privacy of EHRs data and effective clinical decision support system. The proposed work on EHRs retrieval by memory and model-based CF (CBPLSA algorithm – Cluster-Based Probabilistic Latent Semantic Analysis) for DDM aims to minimize computational complexity and to increase EHRs result accuracy. In the next section, problem definition of proposed EHRs retrieval approach is discussed.

PROBLEM DEFINITION

The proposed work of this paper involves framing EHR retrieval, with model-based and memory-based Collaborative Filtering. Among three model-based CF approaches discussed by Koren Y[13], clustering CF along with Probabilistic Latent Semantic Analysis is implemented. As Xiaoyuan Su et al. [14] discussed on hybrid CF, meta-level hybridization technique model learnt from model-based CF on EHRs will be applied over memory- based CF EHRs. Memory-based CF involves EHRs retrieval by framing cluster by CBPLSA algorithm of old patients' EHRs. In model-based CF (new patients EHRs retrieval), clustering model of EHRs which is applied over memory-based CF is formulated. For each disease to be queried corresponding latent words (say for hypertension datasets, latent words are SBP and DBP) will be identified and its probability values will be fixed by analyzing corresponding datasets. Corresponding values of latent words with higher probability will be considered for checking of inclusion of EHRs. The proposed work will be implemented on three publicly available, real-world datasets obtained from Department of Biostatistics, Vanderbilt University. [Table 1] depicts datasets considered, number of patient records and variables along with number of distributed sites considered.

Table 1: EHR Datasets

Datasets	Hypertension	Diabetes	Meningitis
No. of records	381	403	310
No. of variables	05	19	43
No. of distributed sites	08	02	05

PLSA on EHR

Probability of latent words for each disease to be queried will be calculated using the given formulae[15],

P(Di | Q) = P(Di) * P(Q | Di)
where, P(Di) = 1.0
P(Q | Di) = P(Q ∩ Di) / P(Di)

$$P(Q \cap Di) = \sum_{i=1}^{n} P(Li \mid Di) * P(Q \mid Li)$$

where, $P(Q \mid Li) = P(Q \cap Li) / P(Li)$
$P(Li) = 1.0$

In above formulae P(Di | Q) denotes probability of occurrence of keyword Q in corresponding dataset Di of disease queried which is related to probability of dataset considered P(Di) and probability of occurrence of keyword in corresponding datasets P(Q | Di). P(Q | Di) is formulated by calculation of probability of corresponding latent words of disease queried P(Q | Li) and probability of occurrence of latent words in corresponding disease datasets P(Li | Di) (here i, denotes corresponding latent words for disease queried). Here P(Q ∩ Li) represents calculating probability of latent word corresponding to user query. If identified latent word corresponds to Q then P(Q ∩ Li) is 1 else 0. P(Li | Di) represents calculating occurrence of latent word in dataset. If corresponding latent word occurs in dataset then P(Li | Di) is 1 else 0.

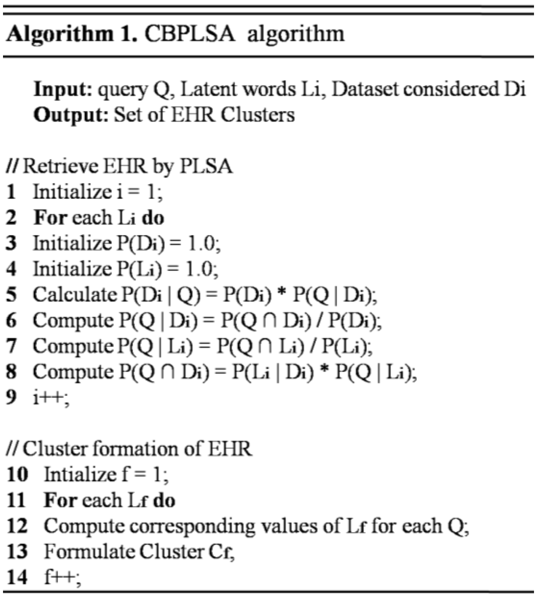
ARCHITECTURAL OVERVIEW

The proposed architectural model on EHR mining involves three-phase operations namely, i) EHR retrieval by memory-based CF ii) EHR retrieval by model-based CF iii) meta-level hybridization technique on retrieved EHR.

In phase I, memory-based CF involves forming cluster of distinct disease identified from history of EHRs by CBPLSA algorithm discussed in algorithm 1. In phase II, retrieval of EHRs by model-based CF (CBPLSA algorithm) on new patients' is done. In phase III, meta-level hybridization technique as stated by Xiaoyuan Su et al.[14] involves model learned from model-based CF (CBPLSA algorithm) being applied on memory-based CF. Further it employs exclusion of redundant EHRs by removing matched patient ID on EHRs retrieved in phase I and phase II.

CBPLSA Algorithm

CBPLSA algorithm is designed with PLSA followed by cluster formation of EHR[15].



Algorithm 1. CBPLSA algorithm

Input: query Q, Latent words Li, Dataset considered Di
Output: Set of EHR Clusters

```
// Retrieve EHR by PLSA
Initialize i = 1;
For each $L_i$ do
Initialize $P(D_i) = 1.0$;
Initialize $P(L_i) = 1.0$;
Calculate $P(D_i \mid Q) = P(D_i) * P(Q \mid D_i)$;
Compute $P(Q \mid D_i) = P(Q \cap D_i) / P(D_i)$;
Compute $P(Q \mid L_i) = P(Q \cap L_i) / P(L_i)$;
Compute $P(Q \cap D_i) = P(L_i \mid D_i) * P(Q \mid L_i)$;
i++;

// Cluster formation of EHR
Intialize f = 1;
For each $L_f$ do
Compute corresponding values of $L_f$ for each Q;
Formulate Cluster $C_f$,
f++;
```

Algorithm 1 of CBPLSA algorithm reveals EHR retrieval by PLSA along with cluster formation. At initial level, for diseases queried, corresponding latent words with higher probability values for each query (disease type queried namely diabetes, hypertension and meningitis) will be identified by PLSA. In next level, those EHR records' with higher or lower values (defined by medical experts) of latent words is formed as cluster (cluster formed within certain range). After that with next latent words values, subsequent cluster will be formed. For disease to be queried, higher probability latent words of corresponding disease have been considered. Based on medical experts' opinion on the abnormal range of corresponding values of latent words (eg: for hypertension datasets latent words are SBP and DBP with normal range of SBP <140 and DBP<90), clusters of EHRs are formulated. For understanding, in-case of hypertension datasets, cluster C1 is EHRs with SBP and DBP in range >290/200, C2 in range <290/200 to >240/150, C3 in range <240/150 to >190/100 and C4 in range <190/100 to >140/90.

EHR retrieval by memory-based CF

In phase I for old patients' record retrieval, memory-based CF is applied. It involves 3-stages namely,

ISSN: 0976-3104

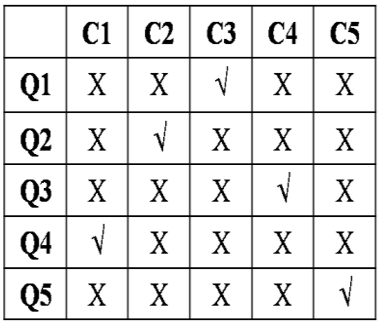
cluster formation by CBPLSA algorithm, matrix formulation and query matching.

Cluster formation by CBPLSA

Clusters for each disease queried have to be formed by CBPLSA algorithm. Latent word analysis, CBPLSA algorithm and cluster formation are the modules for memory-based CF using CBPLSA algorithm.

Matrix formulation

Matrix is formulated [15] with EHRs query Q(example hypertension, diabetes and meningitis) and cluster of EHR records' and is exhibited in [Fig. 1]. A √, tick mark indicates inclusion of corresponding cluster for respective EHRs query and a X, cross mark indicates exclusion of corresponding cluster for respective EHRs query.



	C1	C2	C3	C4	C5
Q1	X	X	√	X	X
Q2	X	√	X	X	X
Q3	X	X	X	√	X
Q4	√	X	X	X	X
Q5	X	X	X	X	√

Fig. 1: Matrix formulation with EHRs query and cluster

Query matching

By memory-based CF, old patients' EHR records' are retrieved by analysing user query Q with formulated matrix. Corresponding clusters computed by CBPLSA algorithm has been retrieved. From [Fig. 1] if query is Q1 (diabetes) among 5 clusters C3 cluster falls under corresponding disease category, similarly for query Q2 (meningitis) among 5 clusters C2 falls under corresponding disease category. Clusters are matched EHRs records'.

EHR retrieval by model-based CF

In phase II, for new patients' record retrieval, model-based CF is applied. CBPLSA algorithm was applied like on-the-fly over new retrieved EHRs. Same phases as EHR retrieval by memory-based CF have to be applied except the difference that clusters formulation to be done using CBPLSA algorithm only in model-based CF.

The major difference between memory-based CF and model-based CF is inclusion of old patients' EHRs for memory-based CF (EHRs retrieval by cluster formation by latent word analysis, matrix formulation and query matching) and new patients' EHRs for model-based CF (EHRs retrieval by latent word analysis and cluster formulation). Memory-based CF involves prediction/retrieval from history/previous preferences/records whereas model-based CF involves prediction/retrieval by on-the-fly approach.

Removing redundant EHRs

In phase III, EHRs retrieved by memory-based CF and model-based CF are accumulated and any redundant EHRs will be excluded for matched patient-ID from both CFs as defined [15]. The final distinctive result of EHRs is corresponding to user query is obtained. To prove its efficiency, this proposed architecture is implemented and their performance is analyzed with state-of-art datasets [Fig. 2].

EXPERIMENTAL IMPLEMENTATION

In this section, implementation detail is presented, followed by description of comparison works and probability values of latent words of EHRs is depicted.

Experimental setup

Experiment was implemented with the proposed architecture using real-world EHRs obtained from Department of Biostatistics, Vanderbilt University. EHR datasets include Hypertension, diabetes and meningitis patient EHRs. Experiment is implemented using C# language and more effective interface is designed which displays latent words along with probability values for disease queried. The results are obtained by implementing three state-of-arts approach of EHR retrieval along with the proposed architecture.

Comparison works

Proposed architecture using CBPLSA algorithm for DDM on EHRs is compared with 3 state-of-arts EHR retrieval approach namely, DDM approach on EHR, memory-based (old patients EHRs) CF on EHR and model-based (new patients EHRs) CF on EHR. Experimental results were analyzed on all the three EHRs datasets considered.

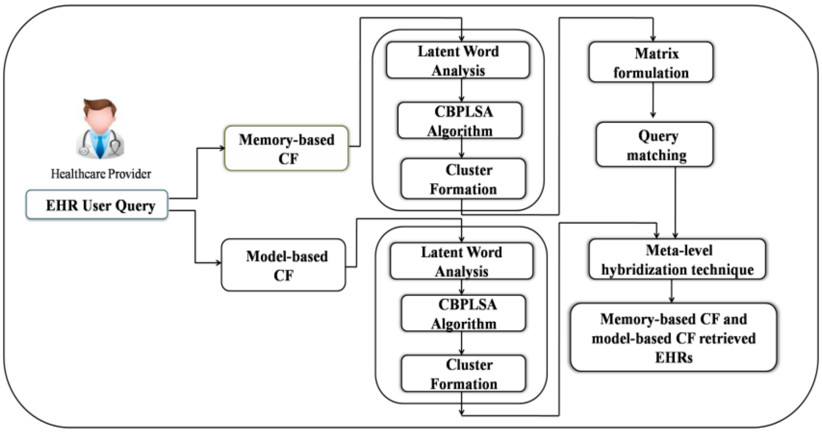


Fig. 2: Architecture - Memory-based and model-based CF (CBPLSA algorithm) for DDM on EHRs retrieval

Performance evaluation

Performance metrics are used for determining the executing capability of proposed program on datasets considered along with computation time required. Evaluation of experimental observations on proposed architecture is analyzed on effectiveness and efficiency measures.

Effectiveness measures

(i) Precision(P): Precision value which is calculated as follows represents a measure that EHRs retrieved is relevant to medical provider query[15].

$$\text{Precision (P)} = \frac{TP}{TP + FP}$$

(ii) Recall(R): Recall value which is calculated as follows represents that relevant EHRs is retrieved by medical provider query[15].

$$\text{Recall (R)} = \frac{TP}{TP + FN}$$

(iii) F-measure: F-measure value which is calculated as follows represents that a higher value of F-measure indicates higher precision and recall values[15].

$$\text{F-measure} = 2 \quad * \quad \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

(iv) Result Accuracy: Accuracy value which is calculated as follows indicates performance of algorithm based on medical provider query[15].

$$\text{Result Accuracy} = \frac{TP + TN}{\text{Total Records}}$$

Efficiency measures

(i) Execution Time: Execution time of proposed algorithm and state-of-arts EHRs retrieval approach are compared by varying the number of EHRs.
(ii) Scalability: Scalability involves calculation of computation time in varying the EHRs datasets. Scalability and execution time goes hand in hand.
(iii) Memory Cost: Memory cost involves computation of memory consumed in running proposed algorithm and state-of-arts EHRs retrieval approach by varying the number of EHRs.

Experiment is implemented with proposed architecture, latent words probability values along with effectiveness and efficiency measures are calculated for 3 EHRs datasets.

Calculation of probability values

Latent words for each disease are fixed based on information obtained from medical experts. Probability values are fixed by PLSA. [Table 2] depicts latent words along with its probability values for each disease. By assigning latent words probability values for each disease, experiment is simulated and performance is analyzed on all the 3 state-of-art approaches along with proposed approach.

Table 2: Latent words and its probability values

Disease	Latent word	Probability
Hypertension	SBP (Systolic Blood Pressure)	0.5389
Hypertension	DBP (diastolic Blood Pressure)	0.6756
Diabetes	chol (Total Cholesterol Level)	0.7645
Diabetes	glyhb (Glycosolated hemoglobin)	0.8956
Diabetes	HDL (High Density Lipoprotein)	0.7845
Meningitis	whites (Total leukocytes)	0.7854
Meningitis	Polys	0.9863

Effectiveness measures observations

[Table 3] depicts the calculated effectiveness measures for the proposed approach along with the comparison works. From [Table 3], the combined approach of memory and model-based CF using CBPLSA algorithm for DDM on EHRs shows higher value of precision, recall and F-measure compared to memory based and model-based CF on EHRs retrieval approaches. The keyword based approach of DDM on EHR shows higher precision, recall and F-measure. As entire dataset is searched sequentially for EHRs retrieval in this case, leads to maximum computational complexity whereas with the clusters formation from history of EHRs, the proposed EHRs retrieval architecture by memory-based CF and model-based CF for DDM has minimized computational complexity and memory overhead.

Table 3: Experimental results. A - keyword based DDM on EHR, B - DDM + memory-based CF on EHR, C - DDM + model-based CF on EHR, D – Proposed CF using CBPLSA for DDM on HER

EHRs Query	Metrics (%)	A	B	C	D
Hyper tension	Precision	1	0.83	0.73	1
	Recall	0.94	0.79	0.54	0.87
	F-measure	0.96	0.80	0.62	0.93
Diabetes	Precision	1	0.91	0.69	1
	Recall	0.95	0.64	0.57	0.93
	F-measure	0.97	0.75	0.62	0.96
Meningitis	Precision	1	0.88	0.82	1
	Recall	0.94	0.61	0.65	0.79
	F-measure	0.96	0.72	0.72	0.88

In [Fig. 3] result accuracy of all the 3 datasets considered along with proposed architecture in-comparison with 3 state-of-arts EHRs retrieval approach is depicted. Approach A has result accuracy value of 85%, approach B of 68%, approach C of 77% and proposed EHRs retrieval by CBPLSA algorithm of DDM of 91%.

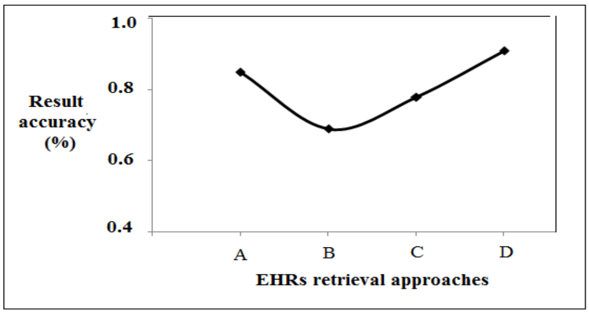


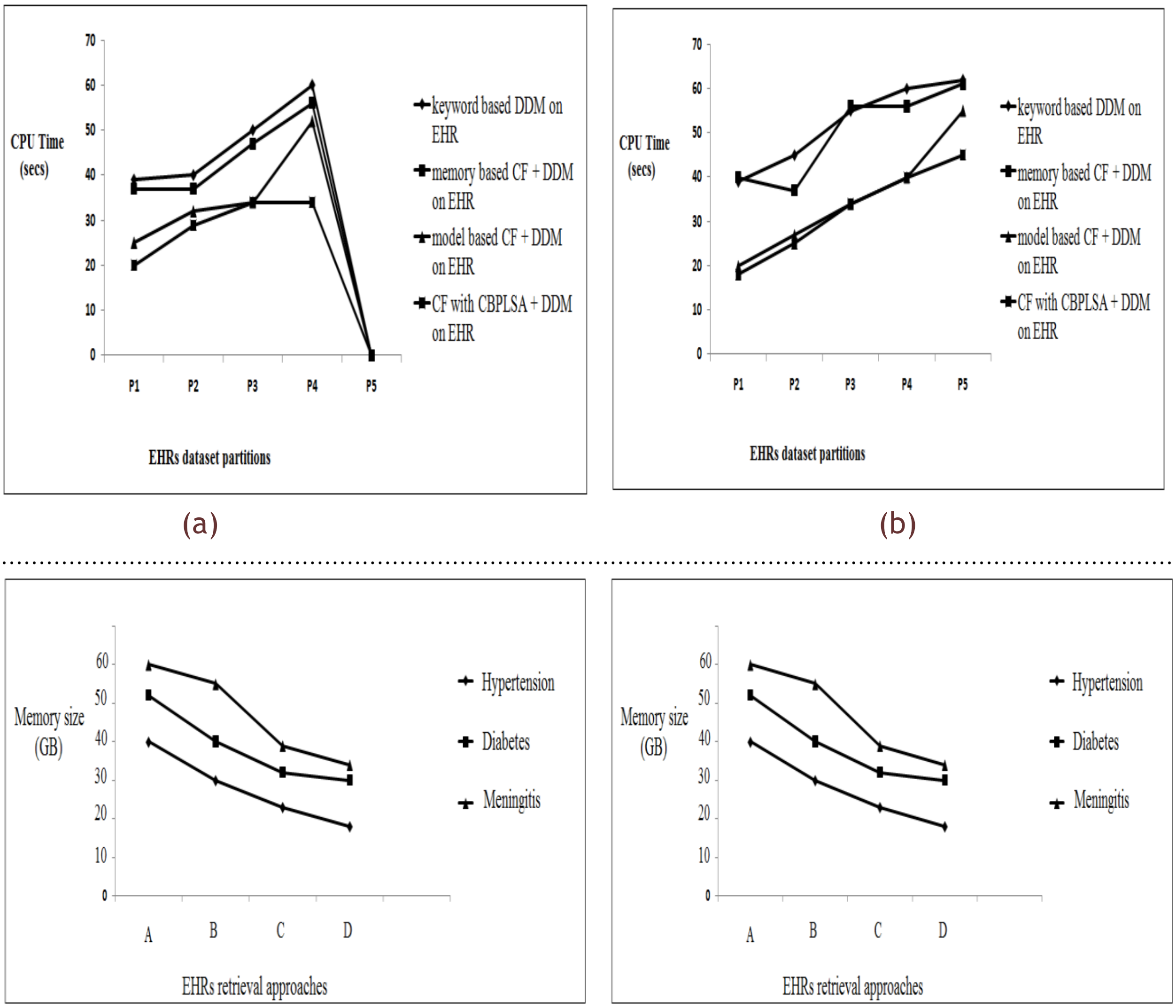
Fig. 3: Result Accuracy of EHRs retrieval approaches.

……

Efficiency measures observations

Efficiency measures namely execution time; memory cost and scalability are calculated by varying the number of EHRs for each transaction. Execution time is measured by varying the number the EHRs with hypertension dataset (D1) of 381 records, diabetes dataset (D2) of 403 records and meningitis dataset (D3) of 310 records. Dividing EHRs into equal partitions (Pi) namely,

P1 – <=100 EHRs
P2 – <=200 EHRs
P3 – <=300 EHRs
P4 – <=400 EHRs
P5 – >400 EHRs



(a)

(b)

(c)

(d)

Fig. 4: Efficiency measures. (a) Dataset D1 vs CPU Time (b) Dataset D2 vs CPU Time (c) Dataset D3 vs CPU Time (d) EHRs retrieval approaches vs Memory size

From [Fig. 4(a), (b) and (c)] it is apparent that execution time for proposed CBPLSA algorithm of DDM on EHRs has minimum CPU runtime in the range of 20 to 30 secs for hypertension EHRs dataset (D1), 21 to 49 secs for diabetes EHRs dataset (D2) and 20 to 48 secs for meningitis EHRs dataset (D3).

From [Fig. 4(d)] it is obvious that memory cost for proposed CBPLSA algorithm of DDM on EHRs increases on increasing the number of EHRs datasets. However compared to other 3 state-of-arts EHRs retrieval approaches, proposed EHRs retrieval by memory-based CF and model-based CF (CBPLSA algorithm) of DDM on EHRs has minimum memory cost on increasing EHRs dataset.

Thus, the proposed architecture of CF using CBPLSA algorithm for DDM on EHRs shows an improved performance in-term of effectiveness and efficiency measures compared to other EHRs retrieval approaches. Further it shows that proposed CBPLSA algorithm minimizes computational complexity.

CONCLUSION

In this paper, architecture for EHR distributed mining with memory-based CF and model-based CF is proposed. Results shown proved that proposed architecture is intended for minimal computational complexity and memory overhead by maintaining clusters of old patients' EHRs. Also, performance evaluation of proposed CF using CBPLSA algorithm for DDM on EHRs shows improved precision, recall, F-measure and result accuracy values than individual memory-based CF on EHRs and model-based CF on EHRs. Further, focuses on finding appropriate CF framework for EHRs retrieval.

CONFLICT OF INTEREST

There is no conflict of interest.

ACKNOWLEDGEMENTS

None

FINANCIAL DISCLOSURE

None

REFERENCES

[1] Sawant V, Shah K. [2013] A review of Distributed Data Mining using agents. International Journal of Advanced Technology & Engineering Research (IJATER), 3(5):27-33.

[2] Cuzzocrea A. [2013] Models and algorithms for high-performance distributed data mining. Elsevier Journal of Parallel and Distributed computing, 73(93):281-283.

[3] Kargupta H, Kamath C, Chan P. [1999] Distributed and Parallel Data Mining: Emergence, Growth and Future Directions. Advances in Distributed Data Mining, (eds.), Hillol Kargupta and Philip Chan, AAAI Press, 407-416.

[4] Zaki MJ, Pan Y. [2002] Introduction: Recent Developments in Parallel and Distributed Data Mining. Springer Journal of Distributed and Parallel Databases, 11(2):123-127.

[5] Park BH, Kargupta H. [2002] Distributed Data Mining: Algorithms, Systems, and Applications. In. Data mining handbook. https://www.csee.umbc.edu/~hillol/PUBS/review.pdf

[6] Tsoumakas G, Vlahavas I. [2008] Distributed Data Mining, Encyclopedia of Data Warehousing and Mining. 2nd Edition John Wang (Ed.), Idea Group Reference, 709-715.

[7] Fu Y. [2001] Distributed Data Mining: An Overview. In: Newsletter of the IEEE Technical Committee on Distributed Processing,5–9.

[8] Khalilia M, Chakraborty S, Popescu M. [2011] Predicting disease risks from highly imbalanced data using random forest. BMC Med Inform Decis Mak. 11:51.

[9] Li Y, Bai C, Reddy CK. [2016] A distributed ensemble approach for mining health care data under privacy constraints. Inf Sci (Ny). 330:245-259..

[10] López-Nores M et al. [2012] Property-based collaborative filtering for health-aware recommender systems. 2011 IEEE International Conference on Consumer Electronics (ICCE). doi: 10.1109/ICCE.2011.5722619

[11] Davis DA et al, [2010] Time to CARE: a collaborative engine for practical disease prediction. Data Min Knowl Disc. 20:388–415.

[12] Zheng K, Mei O, Hanauer DA. [2011] Collaborative search in electronic health records. J Am Med Inform Assoc. 18(3):282-91.

[13] Koren Y, [2008] Tutorial on recent progress in collaborative filtering. In Proc. of the 2nd ACM Conference on Recommender Systems, 8-67.

[14] Su X, Khoshgoftaar TM. [2009] A Survey of Collaborative Filtering Techniques. Advances in Artificial Intelligence. doi:10.1155/2009/421425.

[15] Urmela S, Suresh Joseph K.[2015] An Effective Web Service Selection based on Hybrid Collaborative Filtering and QoS-Trust Evaluation. IJARCET, 3(3):69-78.