Decision Support Systems 88 (2016) 51–66

Contents lists available at ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss

A multi-agent system to support evidence based medicine and clinical decision making via data sharing and data privacy

Hayden Wimmer [a], Victoria Y. Yoon [b,*], Vijayan Sugumaran [c]

[a] Department of Information Technology, Georgia Southern University, P.O. Box 8150, Statesboro, GA 30460, United States
[b] Department of Information Systems, Virginia Commonwealth University, 301 W. Main St., P. O. Box 844000, Richmond, VA 23284-4000, United States
[c] Department of Decision and Information Sciences, School of Business Administration, Oakland University, 442 Elliott Hall, Rochester, MI 48309, United States

ARTICLE INFO

Article history:
Received 19 September 2015
Received in revised form 23 May 2016
Accepted 24 May 2016
Available online 4 June 2016

Keywords:
Multi-agent systems
Privacy preserving data mining
Healthcare informatics
Ontology based data integration
Evidence based medicine
Clinical decision making process

ABSTRACT

Evidence based medicine is the modern standard for clinical decision making where the use of medical evidence combined with clinical expertise and research is leveraged for clinical decisions. Supporting evidence based medicine (EBM) and clinical decision making (CDM) requires access to accurate predictive models and a multidimensional patient view that is aggregated from multiple sources in a multitude of configurations. Data sharing in healthcare remains a challenge due to widespread privacy concerns. Despite abundant research in privacy preserving data mining, healthcare organizations are unwilling to release their medical data on account of the Health Insurance Portability and Accountability Act (HIPAA) requires protecting the confidentiality and security of healthcare data. Further, sensitive data spanning multiple organizations result in not only the data syntax and semantic heterogeneity but also diverse privacy requirements, posing additional challenges to data sharing and integration. In overcoming these challenges, a multi-agent approach is a viable alternative. Despite its potential for addressing the aforementioned issues, little research has been conducted in integrating a multi-agent architecture with privacy preserving data mining in big healthcare data spanning multiple organizations. This research proposes a multi-agent architecture coupled with privacy preserving techniques to facilitate data sharing while preserving privacy. Results indicate that our design artifact is capable of overcoming the aforementioned challenges, thereby facilitating data sharing for knowledge discovery in healthcare and supporting evidence based medicine and clinical decision making via improving predictive models.

1. Introduction

Evidence based medicine (EBM) is becoming the de facto standard in modern medicine. EBM is the use of evidence when making clinical decisions and integrating clinical expertise with available external evidence and research [48]. New strategies are needed to support health information exchange [55]. A complete view of the patient is required to support EBM and clinical decision making. A partial picture, such as missing data attributes or instances prohibits a clinician from making an accurate decision. In diagnostic decisions, misdiagnosis is expensive and potentially life threatening. In the case of small local healthcare organizations (LHO), the patient samples may not be adequate to construct accurate predictive models, thereby making data integration a necessary aspect of EMB. Similarly, physicians and other clinical decision makers require access to multiple predictive models to aid in forming a diagnostic decision. In order to build accurate predictive models and support EBM in the clinical decision making process, data integration from multiple sources is oftentimes a requirement.

Data sharing and integration is a huge challenge given the fact that there is a great explosion in the amount of data being generated [8,12,17]. If done successfully, data integration from multiple sources can lead to productivity gains and competitive advantage. Unfortunately, widespread privacy concerns aggravate the challenge of data sharing and integration for big data analytics [2]. Medical data is highly sensitive that the federal Health Insurance Portability and Accountability Act (HIPAA) requires protecting the confidentiality and security of healthcare data [11]. Security of patient privacy in health information exchanges (HIE) has become a research priority [24]. Privacy preserving data mining (PPDM) has developed various approaches for protecting the privacy of individuals or groups within a dataset while maintaining the integrity of the knowledge contained within the data for knowledge discovery purposes. Despite abundant research in PPDM, healthcare organizations are concerned with potential breach of data privacy and thus unwilling to release their medical data. Further, sensitive data spanning multiple organizations result in not only data syntax and semantic heterogeneity but also diverse privacy requirements, posing additional challenges to data sharing and integration. Overcoming the

* Corresponding author.
E-mail addresses: hwimmer@georgiasouthern.edu (H. Wimmer), vyyoon@vcu.edu (V.Y. Yoon), sugumara@oakland.edu (V. Sugumaran).

http://dx.doi.org/10.1016/j.dss.2016.05.008
0167-9236/

challenges of data sharing and integration in healthcare calls for a broad framework that facilitates customized inter-organizational data exchange, integration, and mining, all based on organizational requirements while maintaining privacy via the de-identification of individual data to adhere to HIPAA standards.

Multi-agent systems perform tasks and take action on behalf of users [39]. The architecture of such multi-agent systems is inherently distributed since they employ multiple agents to perform distinct tasks and operations in different capacities or in different physical or logical locations. Individual agents could be customized to meet the privacy requirements of individual data sources before transmission to a central repository for knowledge discovery. A collection of such agents could facilitate data integration and maintain data privacy as well as the integrity of source data for EBM in the clinical decision making process. Despite its potential for addressing the privacy issues in distributed data, few studies have been conducted in integrating a multi-agent architecture with privacy preserving data mining in big healthcare data distributed over multiple organizations.

Our study aims to develop a multi-agent-based approach for preserving privacy at the data source while integrating multiple data sources for knowledge discovery for EBM. Clinicians are frequently required to use data aggregated from multiple sources to support the clinical decision making process [53]. Consider the following scenario: a healthcare organization has multiple disparate data sources and wishes to share them with other organizations or researchers via a shared clinical data warehouse. Data must be integrated at the source and properly formatted for the destination clinical data warehouse (CDW). Additionally, the integrity of individual privacy must be enforced at the local site prior to sharing the data via the shared CDW. In order to address the integration and privacy issues of this common scenario, this paper presents a design artifact, called Multi-Agent Privacy Preserving for Medical Data (MAPP4MD), which follows the design science research method [21,42]. Our design artifact, MAPP4MD, functions as a necessary precursor to intelligent clinical decision support and evidence based medicine. Specifically, MAPP4MD aids in the clinical decision making process by integrating data from multiple sources so clinicians have access to more accurate predictive models. The results show that MAPP4MD is capable of preserving data privacy at the source prior to integration into larger datasets, thereby overcoming the privacy concerns which accompany integrating medical data into a central data repository for knowledge discovery. Decisions are supported via augmenting their current data with additional attributes and instances to facilitate automatic generation of more accurate predictive models. The improved robustness of the predictive models arms the physician with otherwise unavailable information and models that support the decision and diagnostic process.

The remainder of this paper is organized as follows. Section 2 provides the literature review; Section 3 discusses the conceptual architecture of MAPP4MD and the implementation details. Section 4 presents the evaluation of MAPP4MD followed by the discussion in Section 5. Finally, Section 6 offers conclusions and future research directions.

2. Literature review

2.1. Privacy preservation and data integration

Data anonymization is a key aspect of privacy preservation, particularly in healthcare. Anonymization and Pseudonymity are two important privacy measures which are used while sharing sensitive data [16]. In healthcare, the data is typically horizontally distributed with different sites holding same set of attributes for different individuals. In this case, private information that directly identifies an individual is removed from other medical data. In anonymization, a certain degree of fuzziness is introduced for the attributes that hold high risk of re-identification to a given data item. A privacy criteria such as k-anonymity is popularly used in regards to quasi-identifiers wherein a data set is called k-anonymous if each data item cannot be distinguished as different from the other (k-1) data items [30]. Generalization and suppression are the two main techniques used for introducing fuzziness in the data for the purpose of anonymization. Generalization hierarchies are used to transform categorical and discrete numeric attributes [10]. As an example, generalization of zip-codes can have 6 levels of hierarchies with level-0 showing all the digits of a zip-code while a level-5 would hide all the digits. Such generalization is called suppression.

Research in privacy preserving data mining [19,63] has also proposed several approaches, such as data perturbation, condensation [16], sanitation [61], data distortion, and cryptographic [63]. Data perturbation inserts 'noise' into records, allowing individual record confidentiality [7]. In the condensation approach, data is condensed into multiple groups with a predefined group size. Statistical information such as mean and correlations about each group is maintained. This information is sufficient to preserve the characteristics of the different records [16,37]. Sanitation technique involves removing or modifying items in a database to increase or reduce the support of some frequent item sets in order to hide sensitive patterns without greatly impacting the other patterns [61]. Data distortion modifies the original data so it becomes difficult to estimate the original value of the individual records. Cryptographic methods use the concepts of cryptography in the data mining process in order to facilitate computation of the output model without revealing details of the original input data [63].

While the aforementioned techniques have addressed the privacy concerns to some extent, they assume that data integration has already been done; thus, the fundamental problems underlying privacy preserving data integration have not been addressed. A broad framework that addresses these problems is all the more crucial within the medical domain where there is a great need for data integration to improve patient care and delivery of services.

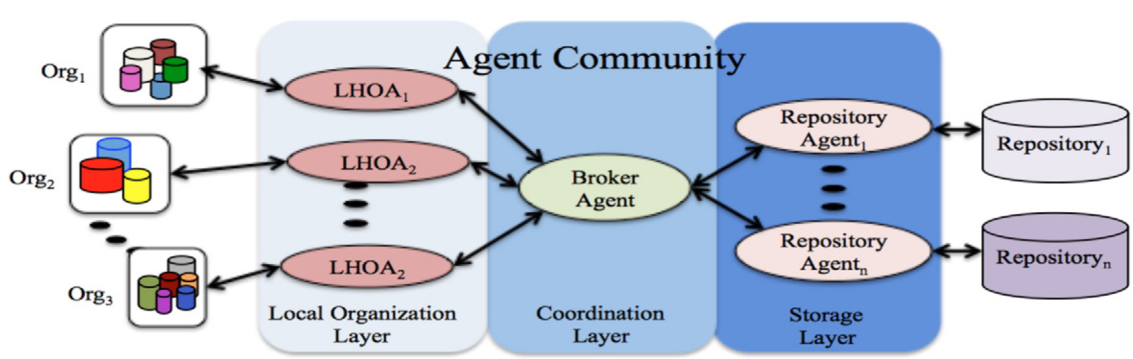

Fig. 1. Architecture of MAPP4MD.