ISSN: 0975-833X

Available online at http://www.journalcra.com

International Journal of Current Research
Vol. 8, Issue, 03, pp. 28293-28297, March, 2016

INTERNATIONAL JOURNAL OF CURRENT RESEARCH

RESEARCH ARTICLE

SURVEY ON DATA MINING METHODS AND APPLICATIONS IN HEALTHCARE DOMAIN SECTOR

***Ramya, R. and Deepika, N.**

Department of Computer Science, NHCE, Bangalore, India

ARTICLE INFO

Article History:
Received 14th December, 2015
Received in revised form
20th January, 2016
Accepted 26th February, 2016
Published online 31st March, 2016

Key words:
Semantic Model,
Knowledge Acquisition,
Different Applications.

ABSTRACT

The successful application of data mining in highly visible fields like e-business, marketing and retail have led to the popularity of its use in knowledge discovery in databases (KDD) in other industries and sectors. Among these sectors that are just discovering data mining are the fields of medicine and public health. This research paper provides a survey of current techniques of KDD, using data mining tools for healthcare and public health. It also discusses critical issues and challenges associated with data mining and healthcare in general. The research found a growing number of data mining applications, including analysis of health care centers for better health policy-making, detection of disease outbreaks and preventable hospital deaths, and detection of fraudulent insurance claims. Electronic health records (EHRs) are representative examples of multimodal/multisource data collections; including measurements, images and free texts. The diversity of such information sources and the increasing amounts of medical data produced by healthcare institutes annually, pose significant challenges in data mining.

Citation: Ramya and Deepika, 2016. "Survey on data mining methods and applications in healthcare domain sector", *International Journal of Current Research*, 8, (03), 28293-28297.

INTRODUCTION

Medical knowledge representation has an exceptional place in the research landscape of medical informatics. The need to unambiguously describe medical knowledge with in clinical environments, inherently characterized by terminological ambiguities, diverse guidelines and data, has given rise to the use of formal ontologies. Semantic models based on such ontologies have been proposed for various medical applications, including computer-aided reporting, medical decision making and data mining. The increasing amounts of medical data produced annually comprise an invaluable source of knowledge to be discovered, represented and exploited to improve health care practices. Data mining, either supervised or unsupervised, provides the methodological tools to extract this knowledge. Supervised methods usually address data classification based on prior knowledge gained by training on previously annotated data, whereas the unsupervised methods group data into clusters based solely on the similarity of the data instances without any training. The latter could be considered as an advantage over the supervised methods. And this study also includes discusses the Data Mining applications in the scientific side.

****Corresponding author: Ramya, R.,***
Department of Computer Science, NHCE, Bangalore, India.

Scientific data mining distinguishes itself in the sense that the nature of the datasets is often very different from traditional market driven data mining applications. In this work, a detailed survey is carried out on data mining applications in the healthcare sector, types of data used and details of the information extracted. Data mining algorithms applied in healthcare industry play a significant role in prediction and diagnosis of the diseases. There are a large number of data mining applications are found in the medical related areas such as Medical device industry, Pharmaceutical Industry and Hospital Management.

To find the useful and hidden knowledge from the database is the purpose behind the application of data mining. Popularly data mining called knowledge discovery from the data. The knowledge discovery is an interactive process, consisting by developing an understanding of the application domain, selecting and creating a data set, preprocessing, data transformation. Data Mining has been used in a variety of applications such as marketing, customer relationship management, engineering, and medicine analysis, expert prediction, web mining and mobile and mobile computing.

MATERIALS AND METHODS

Semantic model

Let us consider a set of medical data acquired from *M* different modalities, with each modality contributing a set of *Ni*, *i*=1...*M*, features within a multimodal data mining environment. For example, such an environment may be defined by the modalities of an intensive care unit (ICU), including a device for the monitoring of the patient's physiological parameters, an x-ray imaging device, and clinicians' free text reports. The physiological measurements, the intensity and textural features extracted from the images, and the textual features extracted from the reports could be considered as feature sets of the respective modalities. In this example these feature sets define four feature spaces, namely the physiological measurements space, the image intensity space, the image texture space, and the textual feature space. The values of each feature set for a particular.

Patient at a particular time instance form a feature vector, represented as a point at the respective feature space; therefore, the patient's status at a time instance is described by four points, each of which is defined at a different space. A cluster of points in a feature space may correspond to different patients or to different time instances of the same patient.

Spatial relations so as to distinguish them from the 2D/3D spatial relations, and they are modeled as concepts. To ensure independency from space dimensionality, the spatial relations can only be defined between 1D projections of a reference and a target object, across a certain axis of a multi-dimensional feature space. Currently two types of spatial relations have been included in our semantic model, namely directional and topological. Each spatial relation can also be linked to its inverse. Directional relations are categorized into positive and negative ones. A positive directional relation represents a direction towards a related object in the feature space and vice versa. Topological relations are divided into eight main categories that are based on region connection calculus 8. The concepts defined in our semantic model are illustrated in Fig. 1(a), and described in the following using DL syntax.

The concept Object refers to a set of objects in a feature space, that are associated through spatial relations between each other. In order to refer to the objects that are used as a reference in the spatial relations, the concept Reference Object has been defined. Target Object refers to objects used as targets in Spatial Relations. The concept Numeric Value, enables the representation of numbers as instances of this concept. This is necessary in order to represent distinct numeric values regardless of their actual value and to overcome the inability of

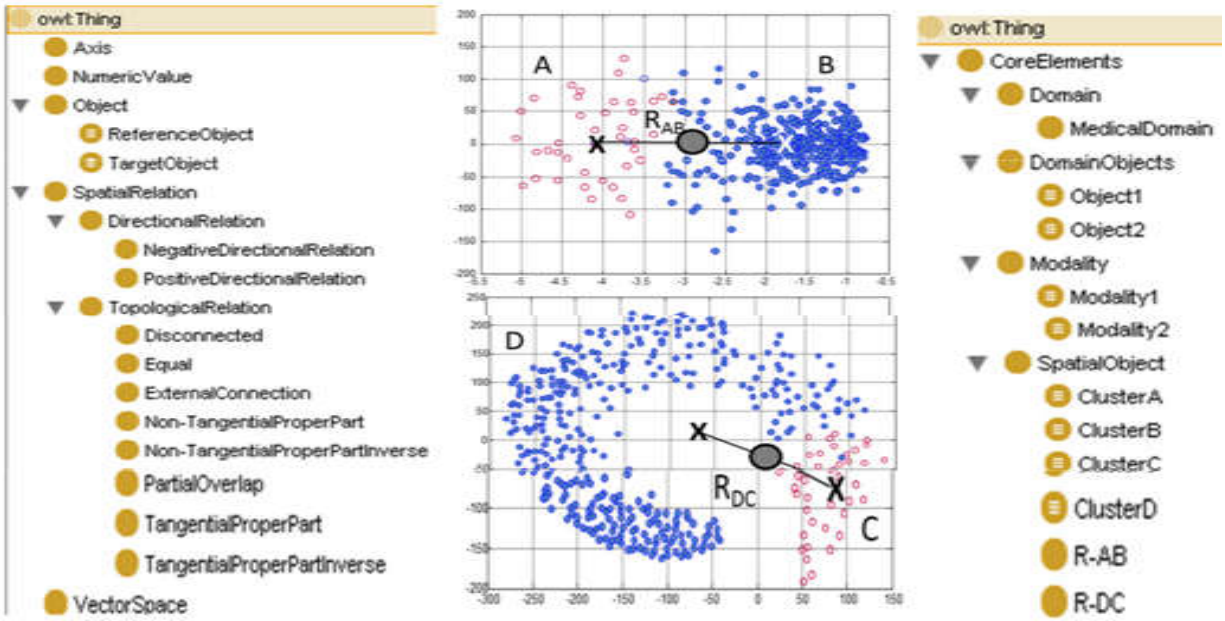

Figure 1. (a) Proposed semantic model. (b) 2D visualization of two example feature spaces from different modalities (Modality 1 and Modality 2). Clusters A and C are formed from samples of Object1, and clusters B and D from samples of Object 2. X-marks represent cluster centroids. (c) Automatically generated ontology from the feature spaces of Fig. 1(b)

However, a reduced representation of the cluster, such as its centroid, may be considered as a simplification in the knowledge representation process. The proposed semantic model enables the representation of knowledge collected by manual annotation of medical data, and of knowledge extracted by data mining, as it can be mapped within feature spaces through the spatial arrangement of objects defined in these spaces, such as points and clusters. These relations are referred to as generalized.

OWL-DL to express numeric data type properties that can be used for reasoning. The concept Vector Space represents a multi-dimensional space. A vector space may be defined by many axes that can also belong to other vector spaces as well: Vector Space⊑ (∃ defined By. Axis) ⊓ (∀ defined By. Axis). The Axis concept represents an axis that may define one or more vector spaces at the same time: Axis ⊑ (∃ defines. Vector Space) ⊓ (∀ defines. Vector Space). Spatial Relation refers to the set of spatial relations defined according to a reference

object and a target object across an Axis: Spatial Relation ⊑ (∃ reference Object) ⊓ (∃ target Object) ⊓ (∃ has Axis. Axis) ⊓ (∃ has Space. Vector Space) ⊓ (∀reference. Object) ⊓ (∀target. Object) ⊓ (∀ has Axis. Axis) ⊓ (∀ has Space. Vector Space) ⊓ (=1 reference) ⊓ (=1 target) ⊓ (=1 has Axis) ⊓ (=1 has Space). The Directional Relation concept refers to the set of relations implying direction across an axis. A Numeric Value indicating the number of intermediate objects (a) (b) (c)

Knowledge acquisition

Given a feature space and a set of annotated objects defined in that space, a new ontology is automatically generated to describe domain knowledge. This is realized by considering the spatial arrangement of the objects in the feature space, using the concepts defined in the previously described semantic model. The generated ontology will have two parts; a fixed part holding fundamental concepts regarding the application domain and the objects, and a dynamically generated part holding the generalized spatial relations between the objects. The fixed part of the concept hierarchy in the automatically generated ontology consists of the class Core Elements, which is superset of all classes in the automatically generated ontology, and four main subclasses of Core Elements (Fig. 1c):

Domain, which represents the user-specified application domain; Domain Objects that contains subclasses of objects that are represented by clusters in each feature space e.g. pathology; Modality, which represents data obtained from a modality; and Spatial Object, which subsumes the automatically generated concepts that represent objects of a feature space e.g. a cluster. In the dynamically generated part, user-specified domains are asserted in the ontology as subclasses of the Domain class. The types of annotated objects are asserted as classes inheriting both the Spatial Object class and a subclass of the domain that represents a user-specified domain. The instances of the annotated objects are asserted as individuals of the class that represents the annotation type. The 1D projection of every object on each axis of the multidimensional feature space is spatially related to the 1Dprojection of a reference object to that axis.

This is realized by means of individuals of the Positive Directional Relation, Negative Directional Relation and Equal. The latter is used to assert that the projections of the two objects are located at the same position on an axis. The reference object can be an arbitrarily selected object. This process is repeated for each modality. The acquired knowledge in the automatically generated ontology can be utilized in a variety of data mining tasks, such as data classification and information retrieval. An example of the automatically generated ontology is illustrated in Fig. 1(c).

DISCUSSION

In order to demonstrate the utility of the proposed model we applied it for classification of anonym zed data obtained from patients hospitalized in ICU. The data include body temperature, blood gasses, and chest x-rays, from which grey-level intensity histogram features and Gabor textural image features have been extracted according to the methodology described in , generating feature spaces as the ones visualized inFig.1(b). The data corresponding to twenty four patients with pneumonia and the image regions corresponding to pneumonia manifestations, known as pulmonary consolidations, have been carefully annotated by clinicians. The domain knowledge was acquired as described in section 1.2 by 10% of the data. This knowledge was used to automatically annotate two unlabeled clusters per feature space, as originating from a sample with pneumonia or not. The clustering was performed by non-negative matrix factorization (NMF). The centroid of the two centroids of the discovered clusters was used as are ference object. The individuals representing the unlabeled clusters were asserted as instances of the class Spatial Object. All classes of the individuals representing the clusters discovered from the rest 90% of the data were successfully inferred by the FACT++ reasoning engine, i.e. all discovered clusters were correctly labeled, regardless of which (disjoint) 10% of the data used for knowledge acquisition.

DATA MINING APPLICATION AREAS

Data mining is driven in part by new applications which require new capabilities that are not currently being supplied by today's technology. These new applications can be naturally into two broad categories.

- Business and E-Commerce
- Scientific, Engineering and Health Care Data

Data Mining Tasks

Data mining tasks are mainly classified into two broad categories:

- Predictive model
- Descriptive model

Medical Device Industry

Healthcare system's one important point is medical device. For best communication work this one is mostly used. Mobile communications and low-cost of wireless bio-sensors have paved the way for development of mobile healthcare applications that supply a convenient, safe and constant way of monitoring of vital signs of patients. Ubiquitous Data Stream Mining (UDM) techniques such as light weight, one-pass data stream mining algorithms can perform real-time analysis on-board small/mobile devices while considering available resources such as battery charge and available memory.

Pharmaceutical Industry

The technology is being used to help the pharmaceutical firms manage their inventories and to develop new product and services. A deep understanding of the knowledge hidden in the Pharma data is vital to a firm's competitive position and organizational decision-making.

Hospital Management

Organizations including modern hospitals are capable of generating and collecting a huge amount of data. Application of

data mining to data stored in a hospital information system in which temporal behavior of global hospital activities is visualized. Three layers of hospital management:

- Services for hospital management
- Services for medical staff
- Services for patients

System Biology

Biological databases contain a wide variety of data types, often with rich relational structure. Consequently multi-relational data mining techniques are frequently applied to biological data. Systems biology is at least as demanding as, and perhaps more demanding than, the genomic challenge that has fired international science and gained public attention.

Treatment effectiveness

Data mining applications can develop to evaluate the effectiveness of medical treatments. Data mining can deliver an analysis of which course of action proves effective by comparing and contrasting causes, symptoms, and courses of treatments.

Healthcare management

Data mining applications can be developed to better identify and track chronic disease states and high-risk patients, design appropriate interventions, and reduce the number of hospital admissions and claims to aid healthcare management. Data mining used to analyze massive volumes of data and statistics to search for patterns that might indicate an attack by bio-terrorists.

Customer relationship management

Customer relationship management is a core approach to managing interactions between commercial organizations-typically banks and retailers-and their customers, it is no less important in a healthcare context. Customer interactions may occur through call centers, physicians' offices, billing departments, inpatient settings, and ambulatory care settings.

Fraud and abuse

Detect fraud and abuses establish norms and then identify unusual or abnormal patterns of claims by physicians, clinics, or others attempt in data mining applications. Data mining applications fraud and abuse applications can highlight inappropriate prescriptions or referrals fraudulent insurance and medical claims.

ISSUES AND CHALLENGES

Applying data mining in the medical field is a very challenging undertaking due to the idiosyncracies of the medical profession. Shilla beer and Roddick's work (2007) cite several inherent conflicts between the traditional methodologies of data mining approaches and medicine. In medical research, data mining starts with a hypothesis and then the results are adjusted to fit the hypothesis. This diverges from standard data mining practise, which simply starts with the data set without an apparent hypothesis. Also, whereas traditional data mining is concerned about patterns and trends in data sets, data mining in medicine is more interested in the minority that do not conform to the patterns and trends.

Table 1. Data Mining Applications In Healthcare

S.No	Type of disease	Data mining tool	Technique	Algorithm	Traditional Method	Accuracy level(%) from DM application
1	Heart Disease	ODND, NCC2	Classification	Naïve	Probability	60
2	Cancer	WEKA	Classification	Rules. Decision Table		97.77
3	HIV/AIDS	WEKA 3.6	Classification, Association Rule Mining	J48	Statistics	81.8
4	Blood Bank Sector	WEKA	Classification	J48		89.9
5	Brain Cancer	K-means Clustering	Clustering	MAFIA		85
6	Tuberculosis	WEKA	Naive Bayes Classifier	KNN	Probability, Statistics	78
7	Diabetes Mellitus	ANN	Classification	C4.5 algorithm	Neural Network	82.6
8	Kidney dialysis	RST	Classification	Decision Making	Statistics	75.97
9	Dengue	SPSS Modeler		C5.0	Statistics	80
10	IVF	ANN, RST	Classification			91
11	Hepatitis C	SNP	Information Gain	Decision rule		73.20

What heightens this difference in approach is the fact that most standard data mining is concerned mostly with describing but not explaining the patterns and trends. In contrast, medicine needs those explanations because a slight difference could change the balance between life or death. For example, anthrax and influenza share the same symptoms of respiratory problems. Lowering the threshold signal in a data mining experiment may either raise an anthrax alarm when there is only a flu outbreak. The converse is even more fatal: a perceived flu outbreak turns out to be an anthrax epidemic (Wong et al 2005). It is no coincidence that we found that, in most of the data mining papers on disease and treatment, the conclusions were almost-always vague and cautious. Many would report encouraging results but recommend further study. This failure to be conclusive indicates the current lack of credibility of data mining in these particular niches of healthcare.

The confusion about the definition of data mining also complicates the issue. For example, we found a couple of papers with the keywords "data mining" in their titles but turned out to be the simple use of graphs. Shillabeer (2009) said that this misunderstanding is prevalent in the relatively young existence of data mining in healthcare. Even if data mining results are credible, convincing the health practitioners to change their habits based on evidence may be a bigger problem. Ayres (2008) reports a couple of cases where hospital doctors refused to change hospital policy even when confronted with evidence. In one case, it was found that doctors coming out of autopsy without washing hands and led to a high probability of deaths in the patients they treated after the autopsy. Presented with this evidence, doctors still refused to change their habits until only much later. Privacy of records and ethical use of patient information is also one big obstacle

for data mining in healthcare. For data mining to be more accurate, it needs a sizeable amount of real records. Healthcare records are private information and yet, using these private records may help stop deadly diseases.

Conclusion

This paper aimed to compare the different data mining application in the healthcare sector for extracting useful information. The prediction of diseases using Data Mining applications is a challenging task but it drastically reduces the human effort and increases the 0diagnostic accuracy. Developing efficient data mining tools for an application could reduce the cost and time constraint in terms of human resources and expertise. Exploring knowledge from the medical data is such a risk task as the data found are noisy, irrelevant and massive too. In this scenario, data mining tools come in handy in exploring of knowledge of the medical data and it is quite interesting. It is observed from this study that a combination of more than one data mining techniques than a single technique for diagnosing or predicting diseases in healthcare sector could yield more promising results. The comparison study shows the interesting results that data mining techniques in all the health care applications give a more encouraging level of accuracy like 97.77% for cancer prediction and around 70 % for estimating the success rate of IVF treatment.

REFERENCES

AdityaSundar, N., PushpaLatha, P. and Rama Chandra, M. 2012. Performance Analysis of Classification Data Mining Techniques over heart disease data base‖, *International Journal of Engineering Science and Advanced Technology*.

Arun K Punjari, 2006. Data Mining Techniques‖, Universities (India) Press Private Limited.

Arvind Sharma and Gupta, P.C. 2012. Predicting the Number of Blood Donors through their Age and Blood Group by using Data Mining Tool‖ *International Journal of Communication and Computer Technologies,* Volume 01 – No.6, Issue: 02 September.

David Page and Mark Craven, Biological Applications of MultiRelationalData Mining‖.

EliasLemuye, ―Hiv Status Predictive Modeling Using Data Mining Technology‖.

HardikManiya, Mosin I. Hasan and Komal P. Patel, ―Comparative study of Naïve Bayes Classifier

Hian Chye Koh and Gerald Tan, Data Mining Applications in Healthcare‖, journal of Healthcare Information Management – Vol 19, No 2.

Jayanthi Ranjan, 2007. Applications of data mining techniques in pharmaceutical industry‖, *Journal of Theoretical and Applied Technology*.

Margaret H. Dunham, 2005. Data Mining Introductory and Advanced Topics‖, Pearson Education (Singapore) Pte. Ltd., India.

Mobile Data Mining for Intelligent Healthcare Support

Prasanna Desikan, Kuo-Wei Hsu, JaideepSrivastava, 2011. Data Mining For Healthcare Management‖, 2011SIAM International Conference on Data Mining, April.

Ruban D. Canlas Jr., MSIT., MBA , Data mining in Healthcare: Current applications and issues‖.

ShusakuTsumoto and Shoji Hirano, Temporal Data Mining in Hospital Information Systems‖.

ShwetaKharya, 2012. Using Data Mining Techniques ForDiagnosis And Prognosis Of Cancer Disease‖, *International Journal of Computer Science, Engineering and Information Technology* (IJCSEIT), Vol.2, No.2, April.

Srinivas, K., Kavitha Rani, B. and Dr. Govrdhan, A. 2010. Applications of Data Mining Techniques in Healthcare and *Prediction of Heart Attacks*‖ *International Journal on Computer Science and Engineering*.

****