A Semantic Model for Multimodal Data Mining in Healthcare Information Systems

M. Rajalakshmi and R. Priya

Abstract--- *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. In this paper we present a novel semantic model that describes knowledge extracted from the lowest-level of a data mining process, where information is represented by multiple features i.e. measurements or numerical descriptors extracted from measurements, images, texts or other medical data, forming multidimensional feature spaces. Knowledge collected by manual annotation or extracted by unsupervised data mining from one or more feature spaces is modeled through generalized qualitative spatial semantics. This model enables a unified representation of knowledge across multimodal data repositories. It contributes to bridging the semantic gap, by enabling direct links between low-level features and higher-level concepts e.g. describing body parts, anatomies and pathological findings. The proposed model has been developed in web ontology language based on description logics (OWL-DL) and can be applied to a variety of data mining tasks in medical informatics. It utility is demonstrated for automatic annotation of medical data.*

I. Introduction

MEDICAL knowledge representation has an exceptional place in the research landscape of medical informatics [1]. The need to unambiguously describe medical knowledge within clinical environments, inherently characterized by terminological ambiguities, diverse guidelines and data, has given rise to the use of formal ontologies [2]. Semantic models based on such ontologies have been proposed for various medical applications, including computer-aided reporting [3], medical decision making and data mining [4].

The increasing amounts of medical data produced annually comprise an invaluable source of knowledge to be discovered, represented and exploited to improve healthcare practices. Data mining, either supervised or unsupervised, provides the methodological tools to extract this knowledge [5]. 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. However, the unsupervised methods, generally, require that the number of target clusters is pre-specified by the user, and the resulting clusters are not associated with class labels.

The data mining approaches described are challenged by the multimodality of the medical data. EHRs include data acquired from multiple modalities such as measurements, images, and free text reports. Methods that have been proposed to cope with this issue include low-level feature fusion [6][7][8] and higher-level ontology-based approaches [9]. *Features* can be measurements, or numerical descriptors extracted from measurements, images, texts or other medical data. Multiple features form multidimensional vector spaces, which are referred to as *feature spaces*. A drawback of the current methods is that the knowledge residing at low-level feature spaces is not directly related to the knowledge represented by higher-level concepts e.g. describing body parts, anatomies and pathological findings. This issue, which is an instance of what is also known as “semantic gap”, can only be partially or indirectly dealt with these methods. In this paper we present a novel semantic model for representation of knowledge extracted from multimodal medical data. It defines generalized qualitative spatial semantics to represent relations between data clusters within multidimensional feature spaces, which in turn can be directly related to higher-level concepts via formal semantics. The proposed model can be considered as a generalization of the model proposed in [10], which was addressing spatial relations between objects only within 2D/3D image spaces, and has enhanced semantic expressivity as compared with our preliminary model presented in [11]. In addition, we present a novel application of the proposed model for automatic annotation of medical data. It is used to provide prior, domain knowledge so that class labels are assigned to the result of a clustering.

II. Methods

Let us consider a set of medical data acquired from M different modalities, with each modality contributing a set of N, 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 patients’

M. Rajalakshmi, Assistant Professor, Department of Computer Science, Sankara College of Science and Commerce. E-mail: mail2rajiravi@gmail.com

R. Priya, Research Scholar, Department of Computer Science, Sankara College of Science and Commerce. E-mail: srpriya.mca@gmail.com

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. However, a reduced representation of the cluster [5], 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 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 (RCC 8) [12]. The concepts defined in our semantic model are illustrated in Fig. 1(a), and described in the following using DL syntax [13]. 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 ReferenceObject has been defined. TargetObject refers to objects used as targets in Spatial Relations. The concept NumericValue, 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 OWL-DL to express numeric datatype properties that can be used for reasoning. The concept VectorSpace represents a multi-dimensional space. A vector space may be defined by many axes that can also belong to other vector spaces as well: VectorSpace ⊑ ($ definedBy.Axis) ⊓ (" definedBy.Axis). The Axis concept represents an axis that may define one or more vector spaces at the same time: Axis ⊑ ($ defines.VectorSpace) ⊓ (" defines.VectorSpace). SpatialRelation refers to the set of spatial relations defined according to a reference object and a target object across an Axis: SpatialRelation ⊑ ($ reference.Object) ⊓ ($ target.Object) ⊓ ($ hasAxis.Axis) ⊓ ($ hasSpace.VectorSpace) ⊓ ("reference.Object) ⊓ ("target.Object) ⊓ (" hasAxis.Axis) ⊓ ("hasSpace.VectorSpace) ⊓ (=1 reference) ⊓ (=1 target) ⊓ (=1 hasAxis) ⊓ (=1 hasSpace). The DirectionalRelation concept refers to the set of relations implying direction across an axis. A NumericValue indicating the number of intermediate objects (or their absence if this value represents zero e.g. Value-0), between the projections of two objects on this axis is required. This way one can uniquely describe the relative position of the target objects in a vector space using a reference object and multiple directional relations:

2.2. 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 CoreElements, which is superset of all classes in the automatically generated ontology, and four main subclasses of CoreElements (Fig. 1c): Domain, which represents the user-specified application domain; DomainObjects 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 SpatialObject, 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 SpatialObject 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 1D projection of a reference object to that axis. This is realized by means of individuals of the PositiveDirectionalRelation, NegativeDirectionalRelation 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.

III. RESULTS AND DISCUSSION

In order to demonstrate the utility of the proposed model we applied it for classification of anonymized 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 [14], generating feature spaces as the ones visualized in Fig.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) [14]. The centroid of the two centroids of the discovered clusters was used as a reference object. The individuals representing the unlabeled clusters were asserted as instances of the class SpatialObject. 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. Future work includes large scale application of the proposed model and its incorporation within our Ratsnake annotation too.

REFERENCES

[1] Rakesh Agrawal, Tomaz Lmielinski and Arun Swami,” Mining association rules between sets of items in large databases”, Proc of ACM SIGMOD Conference on Management of Data, Washington, 1993.

[2] R. Agrawal and R. Srikant, “Mining Sequential Patterns,” Proc. 1995 Int_l Conf. Data Eng. (ICDE _95), pp. 3-14, Mar. 1995.

[3] J.S.Park, M.Chen, P.S.Yu,” An effective hash based algorithm for mining association rules”, Proc. of ACM SIGMOD International Conference on Management of Data, May 1995.

[4] Jiawei Han, Ian Pei and Yiwen Yin,”Mining Patterns without Candidate Generation”, Proc. of 2000, International Conference on Management of Data, May 16-18, 2000 Dallas, Texas, USA.

[5] S. Brin, R. Motwani, and R. Silverstein, “Beyond Market Basket: Generalizing Association Rules to Correlations,” Proc. ACM-1997.