헬스케어 시스템을 위한 세단계 데이터 축소 모델

라하만알리 이승룡, 정태충
경희대학교 컴퓨터 공학과
e-mail : {rahmanali, sylee}@oslab.khu.ac.kr, tcchung@khu.ac.kr

A Three Steps Data Reduction Model for Healthcare Systems

Rahman Ali, Sungyoung Lee and Tae Choong Chung
Dept. of Computer Engineering, Kyung Hee University, Korea

Abstract

In healthcare systems, the accuracy of a classifier for classifying medical diseases depends on a reduced dataset. Key to achieve true classification results is the reduction of data to a set of optimal number of significant features. The initial step towards data reduction is the integration of heterogeneous data sources to a unified reduced dataset which is further reduced by considering the range of values of all the attributes and then finally filtering and dropping out the least significant features from the dataset. This paper proposes a three step data reduction model which plays a vital role in the classification process.

1. Introduction

Data reduction is one of the rapidly emerging fields that have a large number of applications in fields where large datasets are collected and analyzed [1]. Healthcare community is facing challenges in the form of data preprocessing and reduction for better analysis and classification.

A three steps data reduction model is the solution for resolving the problem of data reduction and developing an accurate classifier. Literature provides information about various data reduction techniques in different domains. In healthcare domain, heterogeneous data sources integration is applied to a number of areas such as, molecular biology and hospital information systems [2]. Sokolov et al. [3] have presented a multi-view extension to the GOstruct structured output protein function prediction framework which is capable of combining multiple heterogeneous sources of annotated proteins data from multiple species, and species-specific data. Troyanskaya and his co-authors[4] have addressed the problem of heterogeneous data fusion with Multisource Association of Genes by Integration of Clusters (MAGIC), a general framework that uses formal Bayesian reasoning to integrate heterogeneous types of high-throughput biological data for accurate gene function prediction. Apart from multiple data sources reduction (fusioning), the range of attributes values can also be reduced by using different techniques such as, discernibility matrix [5], rough set for neural network based classifier[6] etc. Similarly, in article [7], CAIM algorithm is proposed for discretization (values range reduction) to maximize the class-attribute interdependence and generate minimal number of discrete intervals. Feature selection is also an important data reduction approach which used to drop the least significant features from the dataset. Selection of the features is done in an automatic way which improves the predictive power of a classifier[8]. Mohamad et al. [9] have proposed a hybrid Genetic Algorithm and Support Vector Machine based algorithm that deals with finding a small subset of informative genes from gene expression microarray data which maximize the classification accuracy.

In this paper, we propose a three steps data reduction model that reduces data to an optimized set of features. The initial input of the model is a set of heterogeneous data sources. Our model not only consider multiple heterogeneous data sources for reduction but also focuses on the reduction of range of values of the continuous value attributes and dropping of the least significant attributes form the dataset.

2. A Three Step Data Reduction Model

The proposed model for data reduction consists of three steps: heterogeneous data sources reduction, attributes values range reduction and attributes dropping as shown in Figure 1. The following sub-sections describe these components in detail.

2.1. Heterogeneous data sources reduction

Usually, a patient medical record is the combination of data from more than one data sources such as, medical images, clinical observations and in some cases from his/her daily activities such as walking, running, jogging, eating and sleeping etc. Let us assume, we have *n* heterogeneous Data Sources such as, $DS_1 \ldots DS_n$ representing a patient condition then combining them to a single Reduced Dataset, *RS,* is represented as follows.

$$RS = DS_1, \ldots, DS_n \qquad (1)$$

RS takes data from *n* data sources and combines them to an integrated dataset. This process is preceded by the development of a unified data model which is in fact a data structure for holding data from different heterogeneous data sources.

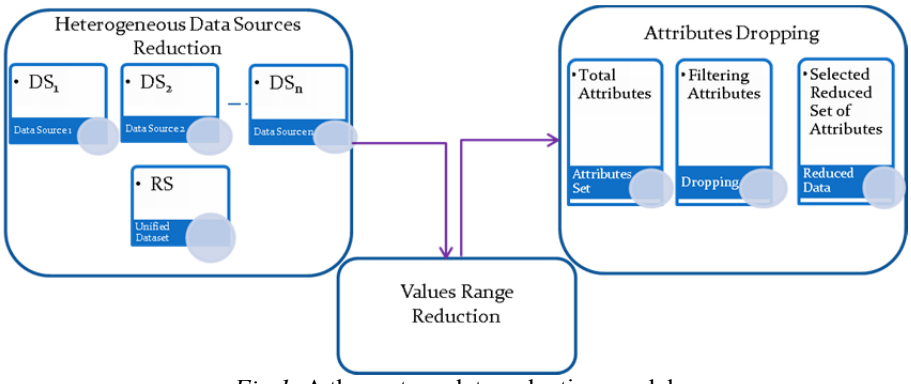

Fig 1: A three steps data reduction model

2.2. Attribute's values range reduction

When once, data is integrated from heterogeneous data sources to a single dataset then the second step is to reduce the range of values of each attribute to discrete intervals. In medical field, the values of attributes are not only discrete but also contain continuous values. These continuous values need to be transformed to discrete intervals so that to reduce its size and support the development of a classifier to accurately classify new queries asked by the patients. Let us consider the range of a continuous value attribute *CA* is from lower value *l* to upper value *u* then it is represented as follows.

$$CA = l \ldots u \qquad (2)$$

In attributes data values reduction, the continuous values are divided into a number of discrete intervals which helps in classification. Consider l_1 is the mid-value of this interval as represented in *eq.3*

$$l_1 = \frac{l+u}{2} \qquad (3)$$

then the discrete values for this attribute will be represented as follows.

$$CA_i = \begin{cases} 0 & if\ CA_i < l_1 \\ 1 & Otherwise \end{cases} \qquad (4)$$

Similarly, if there are more than two values of a continuous value attribute then all the mid-points are calculated by taking average of each two consecutive values. Applying *eq.3* and *eq.4* to each attribute of reduced dataset, *RS* represented by *eq.1,* we get the Discretized Reduced Dataset *DRS* as represented below.

$$DRS = Attributes\ values\ range\ reduction\ (RS) \qquad (5)$$

Eq.5 shows that the input for attribute's values range reduction step is the reduced dataset of the previous step and the corresponding output is the *DRD*.

2.3. Attributes dropping

After reducing the range of values of continuous value attributes, the next step is to reduce the number of features to an optimal set. This is done by identifying the least significant features from *DRS* and dropping them out of the dataset so that we left with the most significant features. The dropped attributes are no more considered for classification. *Eq.6* shows the process of dropping the least significant attributes by taking complement of *DRS* with the significant attributes.

$$Dropped\_attributes = DRS \setminus Significant\_attributes \qquad (6)$$

3. Conclusion

Classification of new instances, as user's queries, in domains such as healthcare requires accurate classifier. The development of such a classifier needs a reduced dataset. The paper has proposed a three steps data reduction model that reduces the data which ultimately helps in the development of an accurate classifier.

Acknowledgement

This work (Grants No. 00048272) was supported by Business for Cooperative R&D between Industry, Academy, and Research Institute funded Korea Small and Medium Business Administration in 2011.

References

[1] Michael, K., *Geometric Data Analysis: An Empirical Approach to d*imensionality Reduction and the Study of Patterns. 2000: John Wiley \\& Sons, Inc. 384.

[2] Sujansky, W., *Heterogeneous database integration in biomedicine.* Journal of biomedical informatics, 2001. 34(4): p. 285-298.

[3] Sokolov, A., et al., *Combining heterogeneous data sources for accurate functional annotation of proteins.* BMC Bioinformatics, 2013. 14(Suppl 3): p. S10.

[4] Troyanskaya, O.G., et al., *A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae).* Proceedings of the National Academy of Sciences, 2003. 100(14): p. 8348-8353.

[5] Gu, J., et al., *A fast way of attributes value reduction.* Acta Scientiarun Naturaltium Universitatis Nankaiensis, 2003. 4: p. 007.

[6] Jelonek, J., K. Krawiec, and R. SlowiÅ„ski, *Rough set reduction of attributes and their domains for neural networks.* Computational Intelligence, 1995. 11(2): p. 339-347.

[7] Kurgan, L.A. and K.J. Cios, *CAIM discretization algorithm.* Knowledge and Data Engineering, IEEE Transactions on, 2004. 16(2): p. 145-153.

[8] Tsang-Hsiang, C., W. Chih-Ping, and V.S. Tseng. *Feature Selection for Medical Data Mining: Comparisons of Expert Judgment and Automatic Approaches*. in *Computer-Based Medical Systems, 2006. CBMS 2006. 19th IEEE International Symposium on*. 2006.

[9] *Mohamad, M.S., S. Deris, and R.M. Illias, a hybrid of genetic algorithm a nd support vector machine for features selection and classification of gen e expression microarray. International Journal of Computational Intellig ence and Applications, 2005. 05(01): p. 91-107.*