Data Collection Methods and Data Pre-processing Techniques for Healthcare Data Using Data Mining

Uma K, M. Hanumanthappa

Abstract— Knowledge Discovery in Databases (KDD) helps organizations and researchers turn their data collection into valuable information. Data collection and pre-processing are the most important and essential stages to acquire the fine and final data that can be taken as correct and suitable for further data mining tasks. Healthcare organizations that take advantage of KDD to lower the healthcare costs while improving healthcare quality by using rapid and better clinical decision making. The data collected by healthcare organization might be structured or unstructured and contains wide variety of data such as hospital details, physician's details, patient records and so on. Healthcare data comprises multimedia content (text, numbers, images, and video). Integrating and transforming these heterogeneous data into single format needs an efficient pre-processing tools. This paper explores the data collection methods and data pre-processing methods and their advantages and disadvantages.

Index Terms— Data collection, Data Cleaning, Data mining, Data preprocessing, Healthcare data, Imputation, Knowledge Discovery in Datbases.

1 Introduction

THE process of finding the new knowledge from the vast amount of historical data is called as Knowledge Discovery in databases (KDD) in other words Data Mining [6]. Data mining is the most motivated area of researchers to discover the meaningful information and finding of patterns in data. The main objective of the data mining is to discover the knowledge hidden in a huge data. Due to rapid growth of clinical data, knowledge mining is becoming a more popular in healthcare industry. Data mining is the most appropriate practice meant for analyzing and discovering the useful information in medical field.

In these days, the data generated at drastic level in every field. Getting the right information from existing data will be most challenging task. Due to abundance of data many academicians and industry researchers are engaged on the process of knowledge mining. Data mining is the core step of Knowledge discovery process. The recent aspects of data availability that are promoting the rapid development of KDD and DM are automatically readiness of data. Data preparation is required to discover the desired knowledge. Hence, the preparation done by various data preprocessing methods and efficient ETL tools to increase the quality of data. Data preprocessing plays a vital role in the KDD process. Data collection method are often loosely controlled, resulting missing values, inconsistence data, noisy data etc. [4, 6]. These type of data can produce ambiguous results in the process of analysis. Hence to generate the accurate result the data should be processed format. Thus, the representation and the quality of data is first and foremost before running the analysis. Data preprocessing includes cleaning, normalization, transformation, feature extraction and selection, etc. The result of data preprocessing is the final training set. Raw data is highly liable to noise, inconsistency and missing values. Data pre-processing is the best solution to improve the quality of data which affects the product of data mining. Data pre-processing is one of the most critical steps in a data mining process which has the concern about preparation and transformation of the initial dataset. Data pre-processing methods are divided into following four category: i) Data Cleaning-Handling Noisy, inconsistency data and missing value in heart disease database. ii) Data Integration- Combining the data from multiple sources such as physician notes, lab records etc. iii) Data Transformation-Consolidation of multiple data formats into a single standard format. iv) Data Reduction- Transformation of masses of data into a small number of summarized reports.

The medical data is often incomplete, inconsistent or lacking in certain many errors. The raw medical data are inadequate to take right decisions sometimes. While the pre-processing techniques results the accurate data. Today, healthcare organization generates a plenty of data which are structured, unstructured and semi structured format. The healthcare data collected from various data sources like hospitals, clinics, doc-

- *Uma K (Research Scholar)*
  Department of Computer Science and Applications
  Bangalore University, Bengaluru, India.
  E-mail: umak@bub.ernet.in
- M. Hanumathappa (Professor)
  Department of Computer Science and Applications
  Bangalore University, Bengaluru, India.
  E-mail: hanu6572@bub.ernet.in

tor's note, patient records and online. The integration of heterogeneous data drawn from various sources can be done using ETL tools or traditional pre-processing methods such as Excel, SQL databases. Thereafter transforming them into a single standardized format also done by numerous existing pre-processing techniques and tools.

2 Knowledge Discovery in Databases (KDD)

The Knowledge discovery process (Figure.1) is an iterative and interactive method consisting of nine steps. The recurrence process at each step, meaning that moving back to previous steps may be required. The procedure starts with determining the KDD goals, and ends with the execution of the discovered knowledge [6].

2.1 Data understanding with application domain: This is the primary preparatory step. It prepares the picture for understanding of data which is necessary for relevance, domain, objectives etc. Having understood the KDD goals, the preprocessing of the data will starts.

2.2 Selecting and creating dataset: After defining the KDD objectives, the data will be used for the knowledge discovery should be identified. This includes evaluation of what data is available, obtaining additional necessary data, and then in combining all the data for the knowledge discovery into one data set, including the property that will be considered for the process.

2.3 Preprocessing and cleaning: In this stage, data reliability is enhanced. It includes data clearing, such as handling missing and inconsistence values, and removal of noisy data or outliers. Among all the steps of KDD process data cleaning plays a vital role in knowledge discovery process. This is because it removes the duplicate records, unnecessary data fields and standardize the data format.

2.4 Data Transformation: In this stage, the generation of better data for the data mining is outfitted and developed. Methods here include attribute transformation and dimensionality reduction. This step is to transforms the data more efficiently based on the desired goal.

2.5 Data Mining Task selection: The transformed data now ready to decide on which type of Data mining to use. An automated search for pattern hidden from a huge data using the selected data mining methods such as classification, clustering, association rule discovery and so on.

2.6 Choosing the Data Mining algorithm: This stage includes choosing the definite method to be used for searching patterns For example in taking into account of accuracy versus understandability, the earlier is better with neural networks, while then present is better with decision trees. This come up to understand the conditions under which a Data Mining algorithm is most suitable. Each algorithm has parameters and strategy of learning.

2.7 Employing the Data Mining algorithm: Finally the performance of the Data Mining algorithm is reached. In this step we might need to take up the algorithm several times until a preferred result is obtained, for instance by modification the algorithm's control parameters, such as the minimum number of time in a single leaf of a decision tree.

2.8 Pattern evaluation: This is the post processing step in KDD process. Identifies the interesting patterns representing the knowledge. Here the preprocessing step is consider after obtaining their outcome on the Data Mining algorithm.

2.9 Using the discovered knowledge: The knowledge discovered ready to incorporate the knowledge into another system for further action. The knowledge becomes active in the sense that can make changes to the system and measure the effects. Actually the success of this step resolves the efficiency of the entire KDD process.

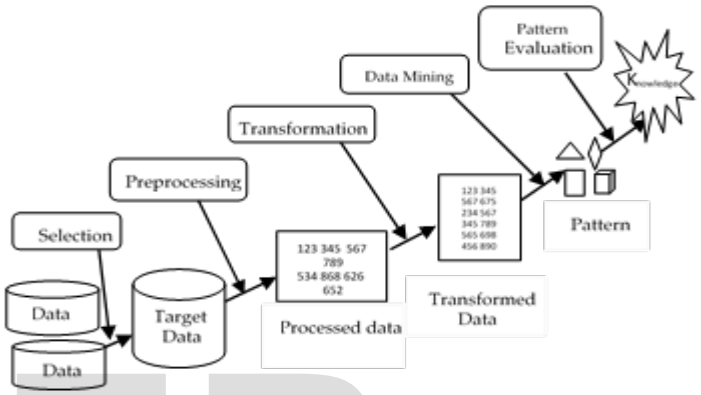

Figure.1- The process of Knowledge Discovery in Databases.

3. Healtcare Data collection methods

Data collection is a process of gathering and determining information on targeted variable is a created methodical way, which then assists one to answer the relevant answer and evaluate the outcomes.

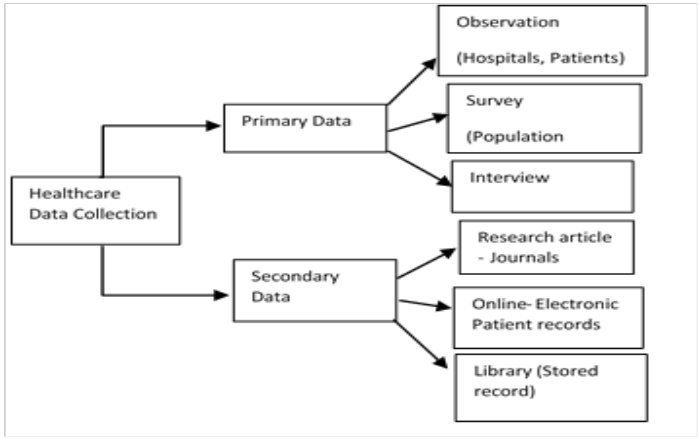

Figure.2. Healthcare Data collection methods

Data can be collected in two data sources, which are categorized as primary data sources and secondary data sources [3].
Primary data: Data that are collected freshly for the first time. Examples- Observation, survey, interview and focus groups.
Secondary data: Data that have been already collected analyzed

by someone else. Examples- Research articles, Internet and library. In healthcare sector an abundant of data generated rapidly which in turn lack of information. There are plenty of methods by which data may be collected for research and healthcare management.

Since healthcare data are heterogeneous in nature data can be collected using both primary and secondary sources. However primary data collection is a very challenging task for researchers. Getting the medical data directly is difficult job since it is confidential and nobody wants to share. Hospitals tend to have information systems for data collection and reporting, staff who are used to collecting registration and admissions data, and an organizational culture that is familiar with the tools of quality improvement, they are relatively well positioned to collect patients' demographic data. In addition, hospitals have a history of collecting race data. The structure and capabilities of primary and specialty care entities vary tremendously, ranging from large groups or health centers with highly structured staff and advanced information systems to solo physician practices with correspondingly small staff. The ability and motivation of these entities to collect and effectively use race, ethnicity, and language data consequently also vary given the investments in Health IT systems and staff training required for these functions. At the same time, these settings have direct contact with patients, ideally as part of an ongoing caregiving relationship. Thus, they are well suited to explaining the reasons for collecting these data, as well as using the data to assess health care needs and patterns of disparities. Physician practices, however, are less likely than hospitals or CHCs to collect race, ethnicity, and language data from patients. And also collected by survey, questioner, interview, online and libraries.

The primary data collection methods are helps to eliminate the subjective bias, whatever information collected is current information and independent to respondent's variable. In other words, the drawbacks of this methods are very expensive, limited information and in certain subjects the respondent's opinion cannot be recorded. In secondary data collection method, data availability is unlimited but reliability and suitability of data requires.

4 Healthcare Data Prerocessing

The set of techniques used prior to the application of a data mining method is named as data pre-processing for data mining [6] and it is known to be one of the most meaningful issues within the famous Knowledge Discovery from Data process [7,8] as shown in Fig. 1. Since data will likely be imperfect, containing inconsistencies and redundancies is not directly applicable for a starting a data mining process. We must also mention the fast growing of data generation rates and their size in business, industrial, academic and science applications. The bigger amounts of data collected require more sophisticated mechanisms to analyse it. Data pre-processing is able to adapt the data to the requirements posed by each data mining algorithm, enabling to process data that would be unfeasible otherwise.Vast amounts of raw healthcare data is existing in the world, data that cannot be directly treated by humans or manual applications. Knowledge and information cannot be easily obtained due to this huge data growth and neither it can be easily understood or automatically extracted. The healthcare knowledge data is key to success of medical decision and treatment. Medical data pre-processing techniques is also follows the typical data pre-process techniques such as data cleaning, integration, transformation and reduction. Handling medical data is a very complicated task since it comprises real world data.

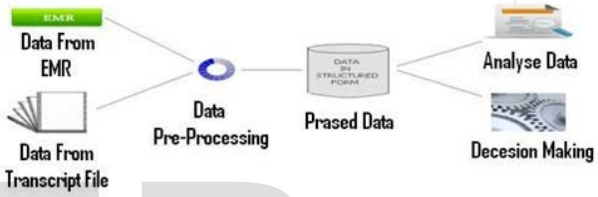

Figure.3. Data pre-processing.

4.1 Data Cleaning

Imperfect data: Most techniques in data mining rely on a data set that is evidently complete or noise-free. However, real-world data is far from being clean or complete. In data pre-processing it is common to employ techniques to either removing the noisy data or to impute (fill in) the missing data. Medical records contains many imperfect data like missing certain fields to fill by patients due to emergency cases, in some context collected data will be noisy. The following two sections are devoted missing values, imputation and noise filtering.

Missing values imputation – Filling the missing value in medical data is very difficult task. Inappropriately handling the missing values will easily lead to poor knowledge extracted and also wrong conclusions [4]. Missing values have been reported to cause loss of efficiency in the knowledge extraction process. Especially in medical field wrong value replacement leads to wrong decision or treatment. To treat the missing values data mining has algorithms such as Expectation-Maximization (EM) algorithm and multiple Imputation algorithm.

Expectation-Maximization Algorithm - is an iterative method to find maximum probability of parameters in statistical models, where the model depends on unobserved latent variables. Given the statistical model which generates a set X

of observed data, a set of unobserved latent data or missing values Z, and a vector of unknown parameters θ, along with a likelihood function

$$L(\theta; X, Z) = p(X, Z \mid \theta),$$

the maximum likelihood estimate (MLE) of the unknown parameters is determined by the marginal likelihood of the observed data [4].

Multiple Imputation - is a statistical method for evaluating incomplete data sets. The work of multiple imputation technique undergoes with three steps: imputation, analysis and pooling.

1. Imputation: Impute or fill in the missing values of the incomplete data sets' m' times. Imputed values are drawn for a distribution. This step results is m complete data sets.
2. Analysis: Analyse each of the *m* completed data sets. This step results in m analyses.
3. Pooling: Integrate the *m* analysis results into a final result. Simple rules exist for combining the m analyses.

Noise treatment - In order to treat noise in data mining, two main approaches are commonly used in the data preprocessing literature. Firstly, corrects the noisy values by using *data polishing methods*, especially it affects the organization of an instance. Even partial noise correction is claimed to be beneficial [4], but it is a difficult task and usually limited to small amounts of noise. The second is to use *noise filters*, which identify and remove the noisy instances in the training data and do no require the data mining technique to be modified.

4. 2 Data Integration

Data integration of medical information to be stored in EMR, HER and/or PHR is a challenging issue. The database heterogeneity problem applies equally to clinical data describing individual patients and biological data characterizing our genome. Specially, databases are highly heterogeneous with respect to the data models they employ, the data schemas they specify, the query languages they support, and the terminologies they recognize. The diverse database systems attempt to merge the different databases by providing uniform conceptual schemas that resolve representational heterogeneities, and by providing querying capabilities that aggregate and integrate distributed data. Research in this area has applied a variety of database and knowledge-based techniques, including semantic data modeling, ontology definition, query translation, query optimization, and terminology mapping.

Figure.4. Data Integration.

4.3 Data Transformation

Data transformation is a process of converting source format data or information into required destination format. Medical data collected from various sources with different formats such as a database file, XML document, or Excel sheet. The first step of data transformation is data mapping. Data mapping determines the relationship between the data elements of two applications and establishes instructions for how the data from the source application is transformed before it is loaded into the target application. In other words, data mapping produces the critical metadata that is needed before the actual data conversion takes place.

Several transformation techniques are applied to transforming source format to targeted format, they are,

- Smoothing: removes noise from data
- Aggregation: summarization, data cube construction
- Generalization: concept hierarchy climbing
- Normalization: scaled to fall within a small, specified range
- Attribute/feature construction

For medical data, all existing transformation methods are not suitable. The data is key to make decision for giving treatment and analysis.

4.4 Dimensionality reduction

When data sets become large in the number of predictor variables or the number of instances, data mining algorithms face the curse of dimensionality problem [6]. It is a serious problem as it will obstruct the operation of most data mining algorithms as the computational cost rise. This phase will emphasize the most influential dimensionality reduction algorithms according to the division established into Feature Selection (FS) and space transformation based methods. Medical data reduction is bit complicated since every instance are plays a important role in some other analysis. The data analysis gives about hospital managemnt, patient diagnosis, and fraud detection and so on.

a) Feature selection: Feature selection (FS) is "the process of identifying and removing as much irrelevant and redundant information as possible" [8]. The goal is to obtain a subset of features from the original problem that still appropriately describe it. This subset is commonly used to train a learner, with added benefits reported in the specialized literature. FS can remove irrelevant and redundant features which may induce accidental correlations in learning algorithms, diminishing their generalization abilities. FS can be used in the data collection stage, saving cost in time, sampling, sensing and personnel used to gather the data.

b) Space transformations: FS is not the only way to cope with the curse of dimensionality by reducing the number of dimensions. Instead of selecting the most promising features, space transformation techniques generate a whole new set of features by combining the original ones. Such a combination can be made obeying different criteria. The first approaches were based on linear methods, as factor analysis [8]

and PCA [9].

c) Instance reduction: A popular approach to minimize the impact of very large data sets in data mining algorithms is the use of Instance Reduction (IR) techniques. They reduce the size of the data set without decreasing the quality of the knowledge that can be extracted from it. Instance reduction is a complementary task regarding FS. It reduces the quantity of data by removing instances or by generating new ones. In the following we describe the most important instance reduction and generation algorithms.

d) Instance selection: Nowadays, instance selection is perceived as necessary. The main problem in instance selection is to identify suitable examples from a very large amount of instances and then prepare them as input for a data mining algorithm. Thus, instance selection is comprised by a series of techniques that must be able to choose a subset of data that can replace the original data set and also being able to fulfill the goal of a data mining application [1, 6]. The original training data by discarding noisy and redundant examples. Instance generation methods, by contrast, besides selecting data, can generate and replace the original data with new artificial data. This process allows it to fill regions in the domain of the problem, which have no representative examples in original data, or to condensate large amounts of instances in less examples. Instance generation methods are often called prototype generation methods, as the artificial examples created tend to act as a representative of a region or a subset of the original instances [5].

4.5. Discretization

It is gaining more and more consideration in the scientific community [16] and it is one of the most used data preprocessing techniques. It transforms quantitative data into qualitative data by dividing the numerical features into a limited number of non-overlapped intervals. Using the boundaries generated, each numerical value is mapped to each interval, thus becoming discrete. Some data mining algorithm that needs nominal data can benefit from discretization methods, since many real-world applications usually produce real valued outputs. For example, three of the ten methods considered as the top ten in data mining need an external or embedded discretization of data: C4.5, Apriori and Naïve Bayes. In these cases, discretization is a crucial previous stage.

5 Related Work

Many researchers have been worked for data collection and pre-processing approach for medical data using traditional methods.

Andrzej walczak et al., [2] proposed a framework to obtain increased accuracy of computer diagnosis in medical for patient check-ups. First researchers created a medical database for asthma, skin allergy and skin illness. The database consists of a descriptive, semantic form of data related as symptom value. All data in the database are created in accordance with the IDC10 standard of illness coding. Then calculating the level of similarity among such common patterns by means of logic operation such as the Jaccard distance. The authors used the method of choosing the shape for single symptoms, say a square and is characterized by means of the following numbers: the area, assumed weight and circumference. So each symptom can be described by those 3 numbers. All equivalent symptoms have numbers equal to 1. Illness pattern parameterization transforms medical database into more divertive forms. Transformation has been done from symptoms expressed in semantic forms to parametric models of symptoms described by set of numbers.

Jing Lu et al., [7] have been worked on the clinical data preprocessing which is use in data mining and analytics. The dataset drawn from Southanpton Breast Cancer Data System (SBCDS) which includes 16,646 breast cancer patients with a total of 23,218 records. The researchers applied data preprocessing techniques to clean the data, to delete or remove the outliers and identify and replace the missing values. The data needs a degree of normalization due to ambiguity of available data. The authors projected the normalization for grouping the patients like i) who are still alive ii) who are have died and cancer has been assigned as the cause of death. iii) Who have died and cancer is the cause of death. After normalization process classification techniques are applied for classify the patient's record to know the survival time.

Ya – Han Hu et al., [9] proposed an approach for large scale medical data pre-processing. The large medical dataset leads the computational cost of the data mining process. The researchers introduced an efficient data pre-processing approach (EDP), which is composed of two steps. The first step is based on training a model over a small amount of training data after performing instance selection. The model is then used to identify the rest of the large amount of training data. For instance selection the authors used three well known algorithms such as IB3, DROP3 and Genetic algorithm. On the other hand, 3 popular classification techniques are used to construct the training models for comparison namely, CART decision tree and K-nearest neighbor (K-NN) and SVM. The experiment is conducted on two datasets namely breast cancer and protein homology. For each dataset, three algorithms are used to generate the 'good' and 'noisy' subsets. Later, classification techniques are applied and measures the performance by means of time complexity. The researchers found that the time complexity of EDP is nearly 2 to 3times less than for the baseline.

6 Conclusion

Huge amount of data generated in healthcare sector which is lacking of potential information to make decision and analysis.

Even though rich data is existing in healthcare, researchers finding difficulties while data collection through primary and secondary data sources. Real world healthcare data contains inconsistency, noisy data which leads to wrong decision and treatment. Healthcare data enforced to preprocess. After data collection, various data preprocessing methods can be applied to clean the data. For handling missing value, imputation method is best for the healthcare data, since every attributes are

plays an important role for decision making. Integrating the medical data needs a very strong knowledge based system and transformation requires mapping of data present in each format. Finally, data reduction is necessary to cut the cost of data management and predict the accurate values from large scale medical data.

REFERENCES

[1] Alastar M. Gray et al., "Applied methods of cost-effectiveness analysis in healthcare" published by OUP Oxford.

[2] Andrzej walczak et al.,"Medical data preprocess for increased selectivity of diagnosis", Bio-Algorithms and Med Systems 2016; 12 (1) : 39 – 43.

[3] C.R. Kothari, Research Methodology: methods and Techniques.

[4] Erhard Rahm et al., "Data Cleaning: Problems and Approaches" University of Leipzig, Germany http://dbs.uni-leipzig.de.

[5] García, S et al., "Big Data Pre-processing: methods and prospects", "Big Data Anal (2016), 1: 9. DOI: 10.1186/s41044-016-0014-0.

[6] Jaiwei Han and Micheline Kamber, "Data Mining Concepts and Techniques", Second Edition, ISBN 13:978-1-55860-901-3.

[7] Jing Lu et al., "Timeline and Episode - Structured clinical data: Pre-processing for data mining and analytics", ICDE 2016 IEEE Workshops.

[8] Usama Fayyad, et al., "From Data Mining to Knowledge Discovery in Databases".

[9] Ya – Han Hu et al.,"An efficient data preprocessing approach for largescale medical data mining", Technology and health care 23 (2015) 153-160, IOS press.