Data Cleaning of Medical Data for Knowledge Mining

Li Jiyun and Wang Junping
Henan Polytechnic, Zhengzhou 450046, Henan, China
Email: leeeteacher@163.com, junping80@163.com

Pei Hongxing*
Zhengzhou University, Zhengzhou 450001, Henan, China
*Corresponding author, Email: phxlee@zzu.edu.cn

***Abstract*—Data mining or data analysis in biomedicine is different from other research fields, because the data in biomedical are heterogeneous and, and they are from different sources. Data from different medical sources are voluminous, each of the resources may have different data structure or data schema, the data quality is also different. Moreover, each physician may have its own interpretation with the same clinical records. In this paper, we analyze the features of medical data, and study data cleaning for medical data in order to mine valuable knowledge. Experiments show that the proposed method is more efficient than the baseline Bayesian network model.**

***Index Terms*—Data Cleaning; Medical Data; Knowledge Mining; Bayesian Network**

I. INTRODUCTION

As the revolution of medical mechanism, Electronic Medical Record has become the main part of medical information media, because it can store a huge amount of data, save resources, is convenient for searching, share information easily, and can also help the enterprises to improve their efficiency of work. These advantages have attracted more attentions from government, enterprises and even the whole society, and make it one of the most important bases in healthcare informatization [1]. However, there are some problems in the development of electronic medical record. The first is the privacy leaking of patients, and this can make a trouble for both the data owner and the patients. For this problem, the institute of medical health should build complete privacy protection architecture to keep patients' privacy, and then study related techniques on the basis of law and rules.

However, the electronic medical record is a valuable treasure for both the research institutes and patients. As the accumulation of patients' electronic medical records, each medical institute has a huge amount of medical data, and these data contain much information about the features of diseases, the clinical diagnosis of patients, and the diffusion of the diseases between people. All these knowledge can be discovered by the data mining techniques. The idea of data mining is that the more data that we have, the more knowledge that we will have [2].

A. Data Mining

Data mining is information process that extracts trusted and efficient knowledge form massive data sources. It aims at extracting unknown but useful knowledge from massive, incomplete, noisy, uncertain and even random data, and solves the problem that traditional statistical method cannot deal with efficiently. The data mining methods mainly include association analysis, classification analysis, clustering analysis, statistical analysis, time-series pattern and deviation analysis. A classical data mining process contains understanding the business, understanding the data, preparing the data, modeling and analyzing, model evaluation and model deployment.

- Understanding the business: We need understand the purposes and requirements of the specific project from the perspective of the hospital, transform them into a definition of data mining, and design a preliminary plan [3]. Evaluate the effect of average length of stay in the hospital to the profit, and design a optimized method, that maximizes the income, and minimize the outlay.
- Understanding the data: After collecting the preliminary data, we need to familiar with them. We need to understand the description of the data, the quality of the data, and so on [4]. In addition, we mainly collect the data about the patients and the diseases, we need to know the relationship between these data, understand concrete meaning of each attribute.
- Preparing the data: The data preparation is preprocessing the raw data in order to construct a dataset that can be processed by modeling tools. It includes the selection of tables, records, attributes, data exchange and data cleaning, and aims mainly at classifying and collecting data from the selected data, and makes them suitable for data mining [5].
- Modeling and analyzing: Apply all kinds of modeling techniques, and optimize the parameters. After the above steps, the data can satisfy the modeling requirements, and thus how to select

doi:10.4304/jnw.8.11.2663-2670

suitable model and process method is critical for the success of the data mining [6].

- Model evaluation: Evaluate the selected model, and check each step of the model to find if the benefit and efficiency are Maximum [7].
- Model deployment: The selection of a good model does not mean the end of the project, and we need to provide the knowledge to users with a convenient manner [8]. Usually, we need to apply the model to the process of decision making. The purpose of data mining is optimization, and in the tradeoff between length of stay and fee, we can achieve a reliable model.

Data warehouse is the preprocessing of raw data before data mining, and it contains structure data that can be easily manipulated according to traditional SQL language. Data cleaning is the process of building the data warehouse. Because the raw biomedical data are voluminous and heterogeneous, and moreover, these data could be gathered from various sources, interview with patients, laboratories, and observations and interpretations from different physicians with their own experiences, so we must preprocess them for further analyzing.

II. Related Work

In this section, we will review related works on knowledge management, data & text mining, and machine Learning and data analysis paradigms in medical informatics.

A. Knowledge Management

Since the MYCIN [9] system was developed in 1970s to support consultation or decision making, many artificial intelligence approaches had been used in data mining in biomedicine domain.

In the MYCIN system, researchers obtain a series of IF-THEN production rules, and this set of production rules is called the knowledge base, and this kind of system is called expert system. In 1980s, the expert systems are very popular. However, with the improvement of modern computers and their cost, more and more knowledge-based medical systems appeared. They adopted the artificial intelligence based expert knowledge as inference rules in the biomedical context, and provided a great opportunity for medical knowledge management, which requires new approaches or technologies in this biomedical field. These systems can not only substitute human diagnosis, but also can be used in the aid of biomedical decision making.

1) Knowledge Management in Health Care

There is a common agreement that the patient record management system is highly developed in medical area [10]. The reason for this is that there is a significant information need for physicians [11], and the amount of data is very large [12]. The biomedical data are mainly text health information about patients, and Hersh classifies them into two main classifications, the health information about the patients and the statistical information from massive information. This kind of classification published in academic journals has been accepted by most researchers. In addition, both of the two types of information are growing rapidly with the development of information technology.

a) Biomedical Knowledge Management

Besides biomedical data, knowledge management has already been used in information retrieval and digital library techniques, where a lot of research articles and reports have been published. In medical data management, the National Library of Medicine (NLM) provides the PubMed service. As we all know that the PubMed service includes millions of biomedical paper citations from MEDLINE and other biomedical institutes.

Moreover, researchers have built a lot of biomedical information retrieval systems, and these systems can help biomedical researchers to find related articles and reports in biomedical paper database systems or on the internet. Data indexing and information retrieval techniques are the most basic and important techniques in information retrieval. Taking the Telemakus system for example [13], it provides an information retrieval framework for medical researchers, and researchers can use it to discovery useful knowledge, and visualize the data mining results. By using information retrieval or visualization techniques, researchers can search the paper database for relevant researchers, and then makes a huge contribution for researchers' significant science finding.

Another example is the HelpfulMed system [14], which can help medical researchers to find related medical documents or reports from medical databases, such as PDQ, MEDLINE, CancerLit, and so on.

2) Accessing Heterogeneous Databases

Although the massive genomic and biomedical data are often distributed in heterogeneous databases, they have made a lot of applications in biomedicine field for all kinds of researchers. In medical data integration, because the biological phenomena are very complex, and the data are hard to explained, we must face with many kinds of challenges whiling integrating these data [15]. In order to share data easily for medical researchers, they have proposed many data integration techniques. For example, Sujansky [16] proposes a heterogeneous database integration framework in biomedicine field, and this framework applies the query-translation technique to provide a uniform conceptual schema for medical data integration. Moreover, the MedBlast system [17] is built to allow researchers to find related articles if given a sequence of keywords.

B. Data & Text Mining

Classical data mining techniques have already been widely used by researchers to find new patterns or new knowledge that have not been acquired in biomedical health field. For example, the Bayesian model is a useful data mining method, and it has already been used in early days. Recently, many new machine learning algorithms, such artificial neural networks and SVM (Support Vector Machines), have been invented and widely used in this area. The above machine learning techniques can be used in different research area, such as proteomics, genomics,

medical diagnosis, and many others. In the following subsections, we will review some of the applications in this research field.

1) Data Mining for Health care

The data mining technique can be used to extract production rules from healthcare data and clinical diagnosis. Researchers have extracted diagnostic rules from survival data using the data mining techniques. [18].

The rules extracted using the data mining techniques are similar with those generated manually by experts. Hence, the results of data mining can be easily validated by domain experts. Moreover, the data mining techniques can also be applied in medical databases, which aim to find new medical knowledge [19]. During all the data mining techniques, classification is the most widely used in biomedical field. Acir et al. [14] applied the SVM model to detect spike signal automatically in Electro Encephalo Grams, and their results can be used in the diagnosis of epilepsy related neurological disorders. Kannan et al. [20] segmented the breast & brain magnetic resonance images using the fuzzy c-means (FCM) algorithm, and their method is effective and automatic.

2) Data Mining for Molecular Biology

The low cost of storage and corresponding sequencing techniques make it very easy to store biomedical data, and these biomedical data can be easily accessed by researchers according the well-developed Internet.

Analyzing these data manually is infeasible for mankind, for there are a lot of annotations and data schemas in these data. And thus, data mining in such kinds of data is necessary, and may be the only solution for this problem. During all kinds of data mining algorithms, clustering [21] and classification [22] algorithms are used extensively in biomedical data.

3) Medical Records Text Mining

Text mining is an efficient data analyzing method, and has been widely used in analyzing medical data.

Because there are a huge amount research papers in public databases, and each research field is quite different from other fields, it is very common to find new sequences or genes. At the same time, the relationships between different biomedical entities are different, so we are not able to know all of them. However, all kinds of medical or gene data are scattered distributed, and if we want to find more knowledge about them, we must link them together [23].

In the whole published literature, all kinds of research fields are included. However, researchers can only specialize in one or some of them, such as several different diseases, so text mining technique is not efficient in the whole database [24]. In order to find some valuable knowledge from the database, we must select a sub-database in only a very small research field, and mine knowledge with domain knowledge.

C. Paradigms of Machine Learning and Data Analysis

Original data analyzing and mining techniques, such as Probabilistic and statistical analysis, have a long history, and they have strong theoretical foundation for data mining and analyzing. The statistical analysis, although not designed for artificial intelligence research, can also achieve similar objectives to machine learning in data analysis and knowledge discovery. Popular statistical analysis techniques, such as multi-dimensional scaling, principal component analysis, time series analysis, discriminant analysis, and regression analysis, have already used widely in medical data, and they are often assumed to be benchmarks for performance comparison between new machine learning techniques with them.

4) Bayesian Model

During all these probabilistic analyzing models, the Bayesian model is most popular, and it has strong theoretical basis for analyzing massive data. Originating in pattern recognition research [25], the Bayesian model classified different objects into predefined classes, and it is based on the features of the objects. Given a class of objects with many features, the Bayesian model keeps the probability of each feature in an object, the probability of each object in a class, and the probability of each class in the whole dataset based on the training sub-dataset. Given a new incoming instance, the Bayesian model classifies it according to the pre-computed probabilities [26]. Based on the original Bayesian model, researchers propose the Naive Bayesian model. The main idea of the Naive Bayesian model assumes that, it assumes a class has many features, but these features are mutually independent. The Naive Bayesian model simplifies the original Bayesian model a lot, and it has been applied in many research fields [27].

Because its simplicity, the Naive Bayesian model has also been widely used in many medical data analyzing fields, such as microarray analysis and genomic.

5) Support Vector Machine Model

In recent years, the SVM (Support Vector Machine) model has gained increasing popularity and recognition in the machine learning researches. The SVM model is also based on the statistical learning theory. The purpose of SVM is to find a hyper-plane, such that the hyper-plane can best divide the data space into several classes [28]. The SVM model has also been used in many researching fields, and many experiments have showed that it is very efficient. In document classification, the SVM model has been proved to have the performance among many machine learning algorithms [29, 30]. In addition, the SVM model have also been used in many medical classification settings, such as physician diagnosing based on clinical records [31] and disease feature classification based on genetic data [27].

III. MEDICAL DATA CLEANING

A. Framework of Data Cleaning

Data cleaning is very important for data analyzing, and it is the preprocessing of dataset before using data mining algorithms, and it is also deeply domain-specific. The data quality problems can either be trivial, or complex or even inconsistent. Currently, there is no common agreement for the definition of data cleaning, because different may have different requirements. The concrete data cleaning process is different from domain to domain, from topic to topic and even from project to project. However, the common understanding for data cleaning is

a process to deal with data of unreasonable, incomplete, and inaccurate. In data cleaning, we can correct the detected errors and add default values to missing features in order to improve the data quality.

The process may include constraint checking, reasonableness checking, completeness checking, and format checking. By reviewing the data, we can identify errors of environmental, temporal, statistical or geographic. We can also evaluate the data quality by expert knowledge of the subject area (e.g. taxonomic specialists).

Figure 1 gives a framework for data cleaning, which is well suited for different purposes. The framework does not contains any data mining algorithm, and it give a guide for data mining researchers to plan the whole data mining process. Some of the data cleaning algorithms will be suitable for the specific work of the data cleaning process. This framework allows the user to interact with the framework by selecting suitable algorithms, but it requires that the user has to know each step of data cleaning clearly, and thus it would be effective in handling noisy data.

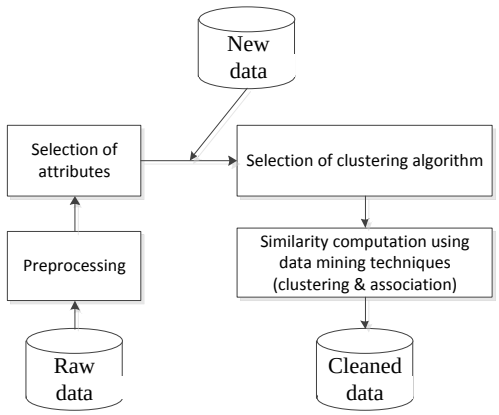

Figure 1. Framework for data cleaning

The principles of the data mining framework are as follows:

- From the raw data, process the data using Pre-processing technique to select suitable attributes;
- Based on the clustering key, group the records by selecting suitable clustering algorithm or blocking method;
- Using data mining techniques, such as clustering algorithm and association rules, mine valuable knowledge by similarity computation.

B. Data Cleaning

Data cleaning is detecting the inconsistencies and errors, and correcting them manually or automatically in order to improve the quality of data. Some business ETL tools have provided some data cleaning function, but they are not expansible. Based on this, some researchers propose the framework of data cleaning system. Data cleaning aims to provide tools for data warehouse or data mining, and then with data warehouse and data mining, we can find patterns or mappings in data set.

Data cleaning for the level instances has many tools, but after pattern transformation or integration, we still need to eliminate the inconsistencies in the instances level. This mainly aims to handle the problems of the missing attributes and instance recognition. During the process of missing attributes, these exists some theories about uncertain information, such as fuzzy set, but they are not suitable for medical, because they need the interferences of users.

However, in order to handle the data cleaning automatically, and solve the duplication problem of data records, we provide an unsupervised marching learning algorithm. The unsupervised marching learning approach can cluster the data records automatically, and handle massive and un-tagged data records. The purpose of clustering algorithm is to generate some clusters, and the records in the same cluster are as similar as possible and in different clusters are as different as possible.

1) Unsupervised Marching Learning

The unsupervised machine learning includes reinforcement learning and competitive learning. Reinforcement learning approach is based on the evaluation, and reinforcing the learning ability from the outer environment. The competitive learning is made up of neutral cells with competition of each other, and these neutral cells compete to get knowledge from the input data. In this paper, based on the classical unsupervised machine learning algorithm, we improve the Hebbian rules, and propose a competitive learning algorithm based on record similarity in order to solve the problem while there are competitions.

According to the Hebbian assumption, we can use a function to represent a rule of a neutral cell:

$$E(w) = -\Phi(w^T x) + \frac{\alpha}{2} \| w \|_2^2 \tag{1}$$

where w is the vector of weight, x is the vector of input specimen, $\Phi(\cdot)$ is a differentiable function, $\alpha \geq 0$ is the forgetting coefficient. The output of the neutral cell is:

$$y = \frac{d\Phi(v)}{dv} = f(v) \tag{2}$$

where $v = w^T x$ is the active coefficient of the neutral cell. According to the gradient descent method, we can have the continuous learning rule:

$$\frac{dw}{dt} = -\mu . \nabla_w E(w) \tag{3}$$

where $\mu > 0$ is the learning speed coefficient, $\nabla_w E(w) = \frac{\partial E(w)}{\partial w}$, so the gradient is:

$$\nabla_w E(w) = -f(v)\frac{\partial v}{\partial w} + \alpha w = -yx + \alpha w \tag{4}$$

and thus, we can get the learning rule of neutral cell as the following:

$$\frac{dw}{dt} = \mu[yx - \alpha w] \tag{5}$$

and the discrete learning rule is

$$w(t+1) = w(t) + \mu\left[y(t+1)x(t+1) - \alpha w(t)\right] \tag{6}$$

If the forgetting coefficient $\alpha = 1$, and the coefficient of rewards and punishments is Y, then when the neutral cell is active (i.e. $y = 1$), the learning rule of the i-th neutral cell is:

$$w_i(t+1) = w_i(t) + \mu Y_i \left[x(t+1) - w_i(t) \right] \tag{7}$$

where

$$Y_i = \begin{cases} 1 & i = k \\ -\beta \dfrac{d_i}{\theta} & i = 1, \ldots, m, \text{ and } i \neq k \end{cases} \tag{8}$$

where β is the punishment coefficient, θ is the threshold of similarity, d_i is the similarity value of the i-th neutral cell, and

$$d_p(x, y) = (\sum_{i=1}^{l} w_i \mid x_i - y_i \mid^p)^{\frac{1}{p}} \tag{9}$$

where x_i and y_i are the i-th elements of vectors x and y, $i = 1, \ldots, l$, and w_i is the weight coefficient.

According to the above description, we have the following unsupervised machine learning algorithm.

(1) Initiate the learning speed coefficient μ, the punishment coefficient β, and the threshold of similarity θ;
(2) Accept the first vector of specimen x, add the first neutral cell w_0, and initiate it as x;
(3) Judge if the learning is over, if so, go to step (5); or else, take an accepting specimen vector from the specimen space, and compute the similarity value according equation (9);
(4) Judge the victory neutral cell by the competitive function,, if $d_i > \theta$, add a new neutral cell and set it as x, and return step (3);
(5) Compute the punishment coefficient by equation (8);
(6) Update the weight by equation (7), and return step (3).

2) Normalized Compression Distance (NCD)

NCD is a distance metrics between two data using the approximation of $K(y)$ where $C(y) = K(y) + k$. The length of the compressed version of y obtained by a lossless compressor C plus the unknown constant k. The NCD is calculated by:

$$NCD(y, x) = \frac{C(y, x) - \min\{C(y), C(x)\}}{\max\{C(y), C(x)\}} \tag{14}$$

where $C(y, x)$ denotes the size of compressed file obtained by the concatenation of y and x.

The NCD represents how different are two files, facilitating the use of this result into various applications into a parameter-free approach. With the patches, we calculate the distance matrix between them using NCD. That is $D = \{d_{ij}\}$, $i, j = 1, \ldots, N$, and N is the number of patches and d_{ij} is the NCD between patch p_i and patch p_j.

IV. Experimental Evaluation

In this section, we evaluated the quality and accuracy of our proposed data clean algorithm over two datasets, a synthetic dataset and a real dataset.

A. Baseline Model

First, we compare our proposed algorithm with the exact Bayesian network model to evaluate the accuracy of the proposed model. In the experiments, our proposed approach consists of two conditional probability distributions (CPDs), $P(I.b \mid M.b, F.b)$ and $P(I.d \mid I.b)$, respectively. To encode the two CPDs, we generate a posterior probability distribution for each of the variables ($I.d$ and $I.b$, respectively) for each probability value of the datasets (i.e., $M.b, F.b$ and $I.b$). Meanwhile, we construct both of the CPDs according to collect instances from the datasets. For example, we can get $P(I.d \mid I.b)$ by the SQL query as follows:

```
SELECT death, birth, count(*)
FROM person
GROUP BY death, birth;
```

With the same reason, we can construct the prior distribution $P(I.b)$ according to the same SQL query. For each constructed CPD, we normalize the raw counts in order to get a probability density by dividing each count of the returned results to sum of all returned counts.

B. Results

First, for each of the experiments, we construct a subset of the original database by copying and altering the raw datasets (synthetic or real dataset). We randomly inject some erroneous information into the subset database, and delete some values to generate some missing information, and after that, we try to find all missing values, make them in the Markov blankets. The reason is that all related items that are not observed would end up in the same blanket. According to the nature of the underlying graphical model, it is very easy to solve these Markov blankets. For example, each item in the Markov blankets is likely to have a birth year, but its death year may be missing. In our experiments, we only consider those blankets with not less than 5 unobserved values. Moreover, in order to minimize the bias introduced by boundary cases, we delete the blankets whose unobserved values are not in the interval between 1850 and 2013.

Next, we learn the parameters of each inference algorithm from the datasets. This can be implemented by the "count group by" SQL queries for the baseline Bayesian model and the "histogram scan" SQL queries in our model. Because our proposed approach is a data-driven algorithm, we return a new data set for this step rather than building a subset dataset from the original dataset.

For fairness, in the experiments upon the corrupted data set, we apply the coarse cleaning method when we construct the model for both of the algorithms. More specifically, we restrict our CPDs of data values including instances between the maximum and the minimum values, i.e., $Min_{DA} \leq I.d - I.b \leq Max_{DA}$. For the

definition of live of a people, we use the ages between 0 and 95, and for any value outside of this interval, we assume that he or she is dead. In addition, we assume the age of a parent is from 18 to 50. The exact inference applies domain knowledge, and that requires the CPDs spanning a certain range of values. However, our approximation inference approach is not based on the limited actual values, and thus has better inference ability.

We ran both algorithms over 100 non-trivial Markov blankets. With this method, we can get results at each level to find the corrupt or missing data, and average them to get a mean value. In this case, most of the Markov blankets are size of 10 to 30. However, there are still some items whose values are as many as 150. For the number of variables is very high and the size of CPD is very large, the exact Bayesian model often ran out of memory. Therefore, we design a test driver to keep the experiments running until we successfully get 100 results.

a) Missing Data

We first consider the effectiveness of both algorithms by varying the level of missing data. In the synthetic dataset, we randomly choose one third of the instances, and delete the birth and death years by setting the value to NULL. This results in several Markov blankets, and then we run the inference algorithms upon them. In the rest of the paper, we abbreviate our algorithm as ERACER, and abbreviate the Bayesian network model as BayesNet.

Because the values of the estimated CPDs $I.b, M.b, F.b$ is very sparse, the BayesNet algorithm has more than twice in the error comparing to our ERACER algorithm. As to the comparison of accuracy between these two algorithms, we implement a shrinkage extension in BayesNet. We use the implemented shrinkage technique to emulate the age-based ERACER algorithm in our setting.

In the BayesNet model, for each blanket of the CPDs, we choose all records with the same age difference, and count them together, instead of approximating probabilities by counting the particular birth/death years, ,for example:

$$P(I.d = 1976 \mid I.b = 1941) = \frac{|\, I.d = 1976 \wedge I.b = 1941 \,|}{|\, I.b = 1941 \,|} \Rightarrow \frac{|\, I.d - I.b = 35 \,|}{N} \quad (15)$$

The results of the experiments are in figure 2. The left two figures of Figure 2 (left) illustrates the accuracy (top) and quality (bottom) for these two data cleaning algorithms, where the Bayesian network model uses the above mentioned shrinkage technique. Both algorithms can get the missing values within the error of three to seven years comparing with the actual values. From the figure wen can see that out proposed ERACER algorithm is more accurate than the BayesNet algorithm, and the reason is that we model the data as a relational dependency network in our algorithm, and the network can add flexibility to the accuracy of the algorithm. However, the BayesNet algorithm is more consistent in terms of quality than our ERACER algorithm. In the BayesNet algorithm, the marginal uncertainty increases along with increasing the amount of missing data. However, the variance of our ERACER algorithm depends on the proposed parameter model, and we use convolution in this experiment.

b) Corrupt Data

Next, we build the experiment based on the precious, but in addition, we introduce some random errors into the synthetic dataset. We corrupt 15% of the individual records randomly, besides deleting 1/3 of them previously. In practice, the mistakes can be either a swapped value, such as 1490 instead of 1940, a missing value, such as 940 instead of 1940, or even a random value between 1850 and 2013. Although our ERACER algorithm can deal with continuous numerical value, it is unnecessary, and the exact inference only requires discrete values in the pre-defined domain. So, we corrupt the selected data by substituting with a random value between 1850 and 2013.

Figure 2 (center) shows the accuracy (top) and quality (bottom) for both of the two algorithms, respectively.

From the figure we can see that the BayesNet algorithm is more resistant to corrupted data values. The reason is that the BayesNet algorithm inferences throughout all Markov blankets jointly. However, the inferred variance of the BayesNet algorithm is higher than our ERACER algorithm, and that is the uncertainty of it is higher than our algorithm because there are contradictory evidences in the dataset. Our algorithm performs nearly as well as the BayesNet algorithm at lower levels of corrupt data, but is more sensitive to the overall error.

The above results are the same as our expectation, for each marginal is independent, and it depends on the voting of the majority instances. In the dataset, when most of the related data is erroneous, the errors can propagate in the defined Markov blanket, and this will make it different to infer the correct data. However, our proposed method is efficient for many applications, because those applications do not have so many errors, and our method can handle them easily. Our proposed algorithm achieves a high accuracy of 95%, which is better than the baseline Bayesian network model, whose accuracy is only 77%. In short, it is obviously that our approximation algorithm performs better that the Bayesian network model in data cleaning.

c) PRF Data

In the experiment of the PRF data set, because the PRF dataset is a real dataset (we do not inject errors in this dataset), and we do not know for certain where the mistakes are, the experimental evaluation is a bit complicated. We ran our data cleaning algorithms on this dataset, and try to recognize the individuals with both birth year and death year. Our algorithm finishes in five rounds, and it identifies those individuals that are not flagged as errors. Then, we generate a test database as before, and inject applicable errors into these seemingly reliable data.

The experiments on the PRF dataset are very insightful, because it contains not only missing or erroneous information of its own, but also contains errors that we

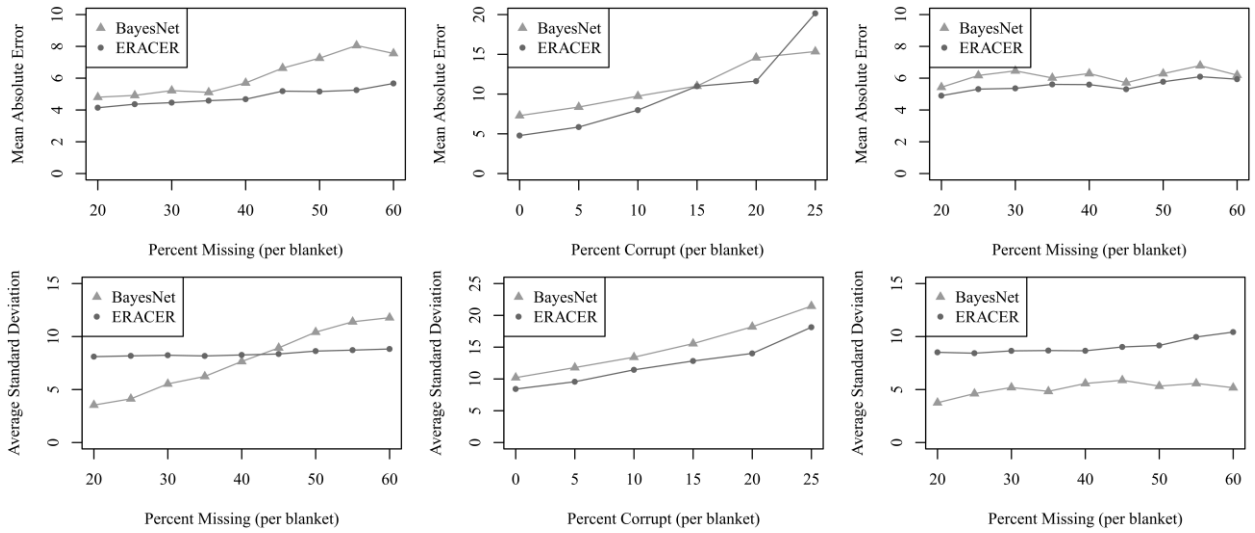

Figure 2. Accuracy (top) and quality (bottom) at varying amounts of missing and corrupt data. Our experiments with the PRF data (right) had results comparable to the synthetic data sets (left and center).

inject manually, and thus this can better illustrate the efficiency of the algorithms. In figure 2 (right), we show the accuracy (top) and quality (bottom) of the two algorithms. From these two figures wen can see that our proposed data cleaning algorithm is more efficient than the Bayesian network model. However, our algorithm will bring a slightly higher level of uncertainty, but this can be ignored in the data cleaning setting.

V. CONCLUSION

In biomedical data analyzing field, the medical data are from different data resources, and thus they would have different data structures or data schemas. Moreover, the interpretations of different physicians to the same clinical records may conclude different results. In addition, to protect the privacy of patients, some personal features many be omitted. In order to mine these massive biomedical data, we must clean them before, and integrate them together to obtain a structured dataset with exact features. In this paper, we analyze the features of medical data, and study data cleaning for biomedical data to mine valuable knowledge. Experiments show that the proposed method is more efficient than the baseline Bayesian network model.

REFERENCES

[1] J. C. Prather, D. F. Lobach, L. K. Goodwin, J. W. Hales and M. L. Hage, et al., Medical data mining: knowledge discovery in a clinical data warehouse., pp. 101, 1997.

[2] S. HettichC. L. Blake and C. J. Merz, UCI repository of machine learning databases, Department of Information and Computer Science, University of California, Irvine, CA, Available at: http//www. ics. uci. edu/\~{} mlearn/MLRepository. html, 1998.

[3] M. Backslash V S Progar, P. Kokol, B. V. S. P. Babi Backslash V, V. Podgorelec and M. Zorman, Vector decision trees, Intelligent data analysis, Vol. 4, No. 3, pp. 305-321, 2000.

[4] K. J. Cios and G. William Moore, Uniqueness of medical data mining, Artificial Intelligence In Medicine, Vol. 26, No. 1, pp. 1-24, 2002.

[5] C. D. Manning and H. Sch U Tze, Foundations of statistical natural language processing, *MIT press*, 1999.

[6] W. Ceusters, Medical natural language understanding as a supporting technology for data mining in healthcare, *Studies In Fuzziness And Soft Computing*, Vol. 60, pp. 41-71, 2001.

[7] G. BrewkaJ. U. R. Dix and K. Konolige, Nonmonotonic reasoning: an overview, CSLI publications Stanford, 1997.

[8] M. Beirlaen and C. Stra Ss Er, Non-monotonic reasoning with normative conflicts in multi-agent deontic logic, *Journal of Logic and Computation*, 2013.

[9] E. H. Shortliffe, Computer-based medical consultations: MYCIN, *Elsevier New York*, 1976.

[10] J. Guptill, Knowledge management in health care, J Health Care Finance, Vol. 31, No. 3, pp. 10-14, 2005.

[11] M. DawesU. Sampson and Others, Knowledge management in clinical practice: a systematic review of information seeking behavior in physicians., *International Journal Of Medical Informatics*, Vol. 71, No. 1, pp. 9, 2003.

[12] A. Sundaram, Information Retrieval: A Health Care Perspective, Bulletin of the Medical Library Association, Vol. 84, No. 4, pp. 591, 1996.

[13] S. S. Fuller, D. Revere, P. F. Bugni and G. M. Martin, A knowledgebase system to enhance scientific discovery: Telemakus, *Biomedical Digital Libraries*, Vol. 1, No. 1, pp. 2, 2004.

[14] H. Chen, A. M. Lally, B. Zhu and M. Chau, HelpfulMed: intelligent searching for medical information over the internet, *Journal Of The American Society For Information Science And Technology*, Vol. 54, No. 7, pp. 683-694, 2003.

[15] J. Barrera, R. M. Cesar-Jr, J. A. O. E. Ferreira and M. D. Gubitoso, An environment for knowledge discovery in biology, Computers In Biology And Medicine, Vol. 34, No. 5, pp. 427-447, 2004.

[16] W. Sujansky, Heterogeneous database integration in biomedicine, *Journal Of Biomedical Informatics*, Vol. 34, No. 4, pp. 285-298, 2001.

[17] Q. TuH. Tang and D. Ding, MedBlast: searching articles related to a biological sequence, *Bioinformatics*, Vol. 20, No. 1, pp. 75-77, 2004.

[18] K. V. R. I. Janurov A and R. Bri V S, A Nonparametric Approach to Medical Survival Data: Uncertainty in the

Context of Risk in Mortality Analysis, Reliability Engineering \& System Safety, 2013.
[19] G. Hripcsak, J. H. Austin, P. O. Alderson and C. Friedman, Use of Natural Language Processing to Translate Clinical Information from a Database of 889, 921 Chest Radiographic Reports1, Radiology, Vol. 224, No. 1, pp. 157-163, 2002.
[20] S. R. Kannan, S. Ramathilagam, R. Devi and E. Hines, Strong fuzzy c-means in medical image data analysis, *Journal Of Systems And Software*, Vol. 85, No. 11, pp. 2425-2438, 2012.
[21] E. C. Wooten and G. S. Huggins, Mind the dbgap: The application of data mining to identify biological mechanisms, Molecular Interventions, Vol. 11, No. 2, pp. 95, 2011.
[22] M. Al Hasan, J. Huan, J. Chen and M. J. Zaki, Biological knowledge discovery and data mining, Scientific Programming, Vol. 20, No. 1, pp. 1-2, 2012.
[23] K. Raza, Application of Data mining in Bioinformatics, arXiv preprint arXiv: 1205. 1125, 2012.
[24] S. Ananiadou, S. Pyysalo, J. Tsujii, D. B. Kell and Others, Event extraction for systems biology by text mining the literature, Trends In Biotechnology, Vol. 28, No. 7, pp. 381-390, 2010.
[25] R. O. DudaP. E. Hart and Others, Pattern classification and scene analysis, Wiley New York, 1973.
[26] N. Mantel and W. Haenszel, Statistical aspects of the analysis of data from retrospective studies of disease, The Challenge of Epidemiology: *Issues and Selected Readings*, Vol. 1, No. 1, pp. 533-553, 2004.
[27] G. I. Webb, J. R. Boughton, F. Zheng, K. M. Ting and H. Salem, Learning by extrapolation from marginal to full-multivariate probability distributions: decreasingly naive Bayesian classification, Machine Learning, Vol. 86, No. 2, pp. 233-272, 2012.
[28] H. Bhavsar and M. H. Panchal, A Review on Support Vector Machine for Data Classification, International Journal of Advanced Research in Computer Engineering \& Technology (IJARCET), Vol. 1, No. 10, pp. 185, 2012.
[29] L. H. Lee, C. H. Wan, R. Rajkumar and D. Isa, An enhanced Support Vector Machine classification framework by using Euclidean distance function for text document categorization, Applied Intelligence, Vol. 37, No. 1, pp. 80-99, 2012.
[30] X. Tang, L. Zhuang, J. Cai and C. Li, Multi-fault classification based on support vector machine trained by chaos particle swarm optimization, Knowledge-Based Systems, Vol. 23, No. 5, pp. 486-490, 2010.
[31] D. Conforti and R. Guido, Kernel based support vector machine via semidefinite programming: Application to medical diagnosis, Computers \& Operations Research, Vol. 37, No. 8, pp. 1389-1394, 2010.

Li Jiyun born in 1974, female, worked at Henan Polytechnic. She is a master and her major is software engineering. Her research area is computer software and application.

Wang Junping born in 1980, female, worked at Henan Polytechnic. She is a master and her major is computer technology. Her research area is computer application technology.

Pei Hongxing born in 1975, male, corresponding author, worked at Zhengzhou University. He is a PHD and his major is mechanical and electronic engineering. His research area is application technology of electrical and mechanical, automation technology, computer application technology.