ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 9, September 2016

Mining Effective Relaxed Functional Dependencies and Association Rules for Electronic Health Data's

Ramesh Kumar B [1], Mohammed Thahmeem T A[2]

Assistant Professor, Department of Computer Science, Sree Narayana Guru College, Coimbatore, Tamil Nadu, India [1]

M.Phil Scholar, Department of Computer Science, Sree Narayana Guru College, Coimbatore, Tamil Nadu, India [2]

ABSTRACT: In the recent scenario, data size is increasing dramatically. So everyone needs to be maintaining different type of dataset based on its functional traits. The functional dependency is the process of generating rules between attributes based on the attribute levels. In this paper, we propose a new functional dependency calculation process for a specific healthcare application, which calculates the dependencies and involved in the drug analysis. In the traditional functional dependency process, there are only few researches implemented with a real time data dependency analysis. In order to perform the real time functional dependency process for drug reaction finding, this paper introduces a new FD method, which is called as RDD (**R**elaxed Functional **D**ependency for **D**rug reaction). This helps to make data cleansing process easier and finally finds the dependency among the set of patient drug consumption attributes. Finally the experiments are carried out using electronic patient health record dataset and the results are obtained from that.

KEYWORDS: Functional dependencies, data quality, Health care analysis, approximate match, Data cleansing.

I. INTRODUCTION

In the recent trend, data is the most indispensable benefit for every domain, which helps to organize and empower the successful business and research area. In a typical data establishment, data is generated in multiple systems and it has become essential to extract meaningful information from the raw data collected from these multiple data sources. These datasets are used for business analysis and decision making process [1]. All successful process on every domain relies on knowledge gathered through examining the information's, performing comparison and correlation between the collected data's. The information management system of an organization normally stores data in databases. The most common type of database is the relational database and forms the information base for the organization. In relational databases, data is stored in tables known as relations. The data in a relation is described by a set of fields known as attributes, denoted by U. The set of values that can be associated with an attribute is known as the attribute domain. Each row of the relation, also known as a tuple, represents an entity that can be described using these attributes. A relational instance r (U) is a set of tuples present in the database table at a particular point of time. A database schema describes how the database is structured and consists of the relations, their attributes, the domain of each attribute and any integrity constraints defined on the relations. The set of constraints included in the schema define the conditions that the data must satisfy.

In the information management, there is a necessity to find the data dependency. The data dependencies have been used to define data integrity constrictions and rules. This helps to improve the quality of database schemas and reduce manipulation irregularities [2]. The data dependency is categorized into 3 types, which are Functional, multi-valued and join dependencies. From the above list, the Functional Dependencies (FD) are the most popular type, this used in database normalization process [3] [4]. This process makes the FD as a popular one.As the relational data model evolved in different directions, also FD theory underwent several extensions to enable the specification of integrity constraints in new application domains. For example, FD were defined for different type of data's one is fuzzy data [5], second one is for XML structured data [6], [7] and finally it supports for multimedia data [8] and temporal data [9]. This paper aims to develop a new dependency analysis for a new type of application domain known as drug

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 9, September 2016

analysis. Medicines are frequently prescribed to patients with the aim of improving each patient's medical state, but an unfortunate consequence of most prescription Medicines is the occurrence of undesirable reactions. Reactions that occur in more than one in a thousand patients are likely to be signaled efficiently by current Medicine surveillance methods, however, these same methods may take decades before generating signals for rarer reactions, risking medical morbidity or mortality in patients prescribed the Medicine while the rare side effect is undiscovered. The system proposes an innovative data mining framework and applies it to mine the dependencies between medical dataset to potential detect associations between various functionalities, where the Medicine related events occur infrequently and abnormally. The proposed system analyzed the Medicine and reactions when it is consumed with other based on its dependency value, when it is related with other medicine and its symptoms based on associations and dependencies. The system aims to propose a data mining algorithm to mine the RDD (Relaxed function for Drug Dependency) signal pairs from electronic patient database based on the new measure.

II. RELATED WORKS

The functional dependencies can be used in various domains and applications such as query relaxation, data integration and analysis, data cleaning, and record matching. And this also used with various domain datasets known as Health care information's, employee dataset, educational dataset and insurance domains. In this paper we have produced the dependency analysis in healthcare dataset, which contain a set of patient electronic health records. The outcome of the application should be an useful one, so this paper aim to create a complete analysis of patient electronic health record and their disease drug lists. The functional dependencies are categorized into several types, which are listed below.

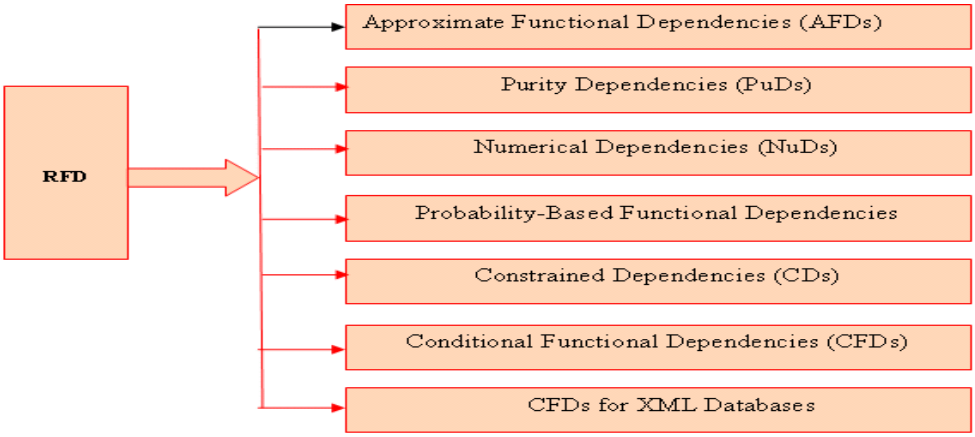

Fig 1.0 Types of RFD

1. **Approximate Functional Dependencies (AFDs):** Approximate Functional Dependencies (AFD) mined from database relations represent potentially interesting patterns and have proven to be useful for various tasks like feature selection, improves classification process, query optimization and query regeneration. The detection of Functional Dependencies (FDs) from a relational database is an unsolved research problem [10]. The AFD suffers from a set of problems such as determining right interesting measures, effective pruning strategy detection and effective traversal process.For example, in the medical database unlikely to have two patients with the same name those have been admitted to the hospital. Thus, except for few homonymy cases, the Name of the Patient should imply the BloodType. Thus, the following AFD might hold:
   D_{true} : $Name_{EQ}$ → $BloodType_{EQ}$:
2. **Purity Dependencies (PuDs):** authors in [11] introduce purity dependencies as generalizations of functional dependencies in relational databases. This initiated from the notion of impurity computation. Finally, an algorithm that mines datasets for these dependencies is presented that is named as Purity Dependencies and Approximate Classifications.
3. **Numerical Dependencies (NuDs):** Numerical dependencies are FDs relaxing on the extent by means of a cardinality constraint. Authors in [12] present a finite set of inference rules for numerical dependencies, which is a

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 9, September 2016

generalization of the existing Armstrong axioms methods. This paper proves that this set is sound and complete for some special cases.

4. **Probability-Based Functional Dependencies:** This type of FD is based on the probabilistic values. In [13] authors propose a framework based on functional dependencies (FDs). Unlike in traditional database design, where FDs are specified as statements of truth about all possible instances of the database; in web environment, FDs are not specified over the data tables. Instead, it generated FDs by counting-based algorithms over many data sources, and extends the FDs with probabilities to capture the inherent uncertainties in them. Given these probabilistic FDs, they have considered two problems to improve data and schema quality in a pay-as-you-go system: (1) pinpointing dirty data sources and (2) normalizing large mediated schemas.
5. **Constrained Dependencies (CDs):** To solving the implication problem for FDs to these dependencies, conditional dependencies are generated. For a large class of constraint domains, satisfying the independence of negative constraints property is a main aim, the approach in [14] provides an efficient solution to the CFD implication problem. For constrained finiteness dependencies we have similar results, although for a class of constraint domains it remains open whether the CFinD implication problem is tractable.
6. **Conditional Functional Dependencies (CFDs):** In [15] authors propose a class of constraints, referred to as conditional functional dependencies (CFDs), and study their applications in data cleaning. The traditional functional dependencies (FDs) that were developed mainly for schema design, but the CFDs aim at capturing the consistency of data by incorporating bindings of semantically related values. In this paper a set of techniques are created for detecting CFD violations.

III. PROPOSED SYSTEM

In this paper, the technique proposes an enhanced RDD for handling data dependencies in patient record and finding the abnormal reactions when the dependencies are changed. This also aimed at performing fully relaxed constraints for dependency calculation and also matches data in different types. The main goal of this work is developing the necessary algorithm and deploying it to an improved health care application. The current work proposes a new RFD model for patient health care domain, which aimed at performing a fine output and summary generation from the dependency score.

A. FUNCTIONAL DEPENDENCY THEORY

The theoretical background of the proposed work is presented in this chapter.

Functional Dependency: A functional dependency (FD) is a constraint between two sets of attributes X and Y in a relation U formally, a function dependency X→Y is satisfied by an instance r(U), when

∀ Pairs of tuples ti, tj ∈ r(U) , if ti[X] = tj[X] , then ti[Y] = tj[Y], X is called the determinant attribute or LHS attribute and Y is called the dependent attribute or RHS attribute.

B. CANDIDATE GENERATE-AND -TEST METHOD

The candidate Generate-and-Test method combines the level-wise search used in apriori association rule mining algorithm with pruning rules to discover FDs from the data. Various algorithms based on the Candidate generate-and-test method that uses attribute partition and level-wise search have been proposed. The proposed method uses a candidate generate-and-test approach by following a level-wise seek through the attribute patterns and at each level the entropy of data's is compared to check for the presence of FDs. In the proposed system the dependencies between drug consumption is analyzed and found the relationship score. The proposed research work aims in extracting dependency score and its conditional outcome for each rule at every iteration. For example consider we have two set of patient records, which need to be integrated and we need to find the dependencies between two records. The dependency score calculation just list out, whether the two records are similar or not. But the score should also give the common traits between the records and rare data from the records. This process is going to be proposed in the system.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 9, September 2016

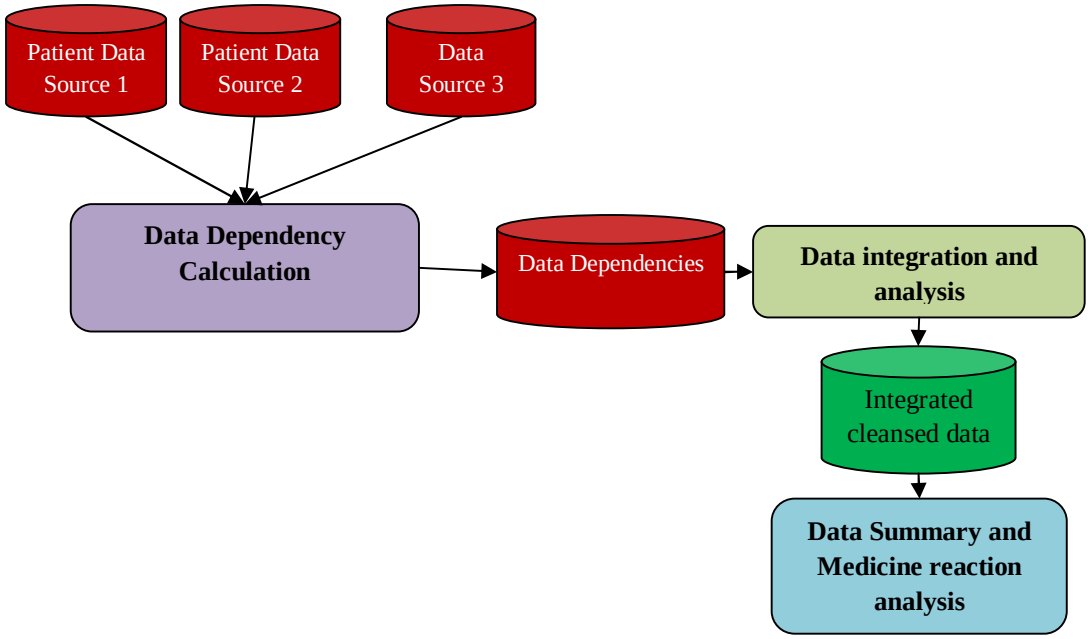

Fig 2.0 Architecture of proposed RDD

Figure 2.0 shows the architecture of a data integration system that uses rule constraints in the form of data dependencies to improve the performance of schema matching, data cleaning and traits comparison operations. The proposed RDD extracts different forms of dependencies from various medical data sources. The data sources are assumed to possess data from similar categories. The discovered dependencies are stored as rules in a repository and are then used to improve the accuracy of schema matching, data cleaning and entity matching tasks in a data integration system. The initial output of the data integration system is the clean integrated data. After the successful integration, the system finds the medicine reaction summary and suggest the optimal and irregular traits. Schema matching is followed by attribute matching to remove duplicate entities from the integrated patient health data.

The collected patient database and its attributes are defined using a hierarchy of data dependencies are used as entity matching rules. In the proposed system, Candidate Generate-and -Test Method has been proposed. It is a form of association rule learning, which seeks to identify meaningful differences between separate drug groups by dependency values and the key predictors that identify for each particular group. This finds the different support valued dataset form the database. For example if “drugA” has high confidence than other drugs, this results as “drugA” is a contrast set group medicine. Take another example given a set of attributes for a pool of patients (labeled by different drug type), a test method would identify the distinct or contrasting features between patients treatment and experienced by some reactions over it. This algorithm helps to discover that the major difference between different drugs related details. The association finding module helps to track all the support and confidence of the drugs. This initially contains the following process. Initially this process performs the following.

- Scan the database and
- Candidate generation
- Support and confidence calculation and
- Generate-and -Test

The frequency detection is the process of identifying the total number of occurrences of prescriptions with various factors. For example, the drug prescriptions for every patient and causal and other symptoms will be captured at every time interval should be identified by the following formula (1).

$$\text{Support } (X)=(T(x))/n \qquad (1)$$

Where x is an attribute, T(x) is the total number of occurrences of x and n is the total transaction in the dataset. In the proposed system the prescriptions data’s are collected and applied for frequency detection. In order to identify the frequency different attributes are considered and applied into the formula (1), thus the final output will be applied

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 9, September 2016

into the association data. As like the support additionally, the confidence among different attributes is considered by the following formula (2).

Confidence (X➔Y)=P(Y/X) (2)

Where x is an attribute, Y) is the total number of occurrences of Y and X is the total number of occurrences of X is the total transaction in the dataset. In the proposed system the prescriptions data's are collected and applied for frequency detection. The rule generation from the confidence is important to analyze the association between attributes. X➔ Y holds with confidence c if c% of the transactions in D that contain X also contain Y. Rules greater than a user specified confidence is calculated to have minimum confidence from the dataset.

T id	Attributes(drug, symptom, other attributes)	Given X➔ Y Confidence=Occurrences of (Y)/ Occurrences of (X)

Table 1.0 confidence calculation of drug type

After the successful calculation of the support and confidence, the rules are filtered based on the support and confidence threshold. Given the items at each level, the dataset is scanned and the support is counted for each group. Each node is then investigated to determine if it is significant and large if it should be pruned and if new leaf should be generated. After all significant contrast sets are located and a post processor selects a subset to show to the user the low order, simpler results are shown first, followed by the higher order results which are unique and significantly different

.

IV. IMPLMENTATION

This section describes the implementation process. Implementation is the realization of an application, or execution of plan, idea, model, design of a research. This section explains the software, datasets and modules which are used to develop the research.

A. *SOFTWARE*: The experiments are performed on an Intel Dual Core with a RAM capacity 2GB. The algorithms are implemented in C#.net and are run under Windows family.

Operating System	**Windows 10**
Front End	**C#.Net**
Back End	**MS SQL server**

Table 2.0 software requirements

B. *DATASETS*: The system used a synthetic dataset, which can be any number of patients created by the proposed data collection phase. Data collection is the first step of the proposed system. The followings are the descriptions about the dataset used for the experiments.

Dataset information's	**values**
Total number Health records(Tuples) record 1	**500**
Total number tuples in record2	**300**
Total number of attributes	**35**

Table 3.0dataset descriptions

The above table 3.0 describes the dataset description used for the experiments. Where the dataset contains 10 number of drug types which are described below.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 9, September 2016

C. ***RESULTS*****:** From the above set of drugs, 128 patients are used drug type1, 141 patients are used drug type2, 60 patients are used drug type3,33 patients are used drug type4 respectively. The details are shown in the following table. The table has been generated from the above dataset D.

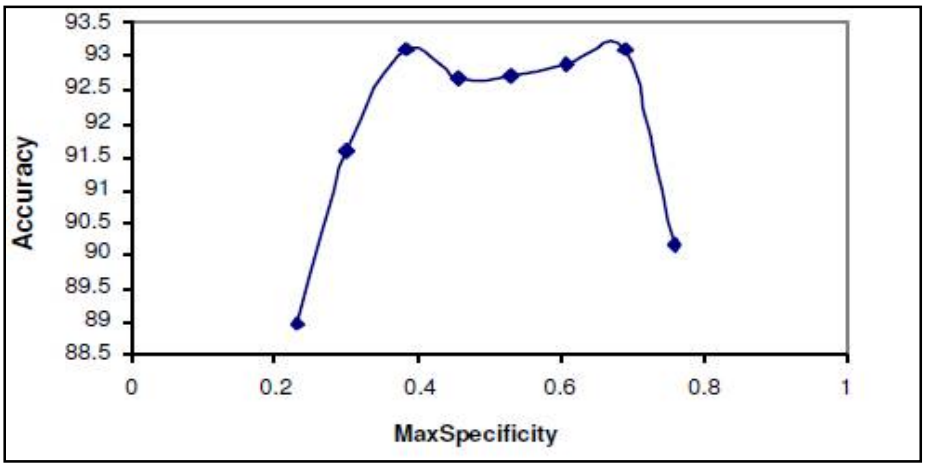

Fig 3.0 accuracy analysis chart

The fig 3.0 shows the plotted the graphs varying the maxSpecificity thresholds for various run of RDD Miner. This is performed over the drug db from patient electronic health record. It represents the classification accuracy of the proposed system based on the specificity function.

V. CONCLUSION

This paper dealt with the management of functional dependency needed databases, and analysis has been performed on it. This paper proposed new FD techniques RDD, which helps to perform record matching and dependency score calculation process. Using RDD, it follows specificity function along with the relaxed functional dependency. This initially performs the data cleansing and finally performs the dependency analysis. At last the system provides an effective analyzed report regarding the medicine reaction based on the different attribute values.

REFERENCES

[1]. Gramatikov, Martin. "Data mining techniques and the decision making process in the bulgarian public administration." *NISP Acee Concerence, Bucharest, Romania* (2003).
[2]. Chu, Fang, et al. "Improving mining quality by exploiting data dependency."*Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer Berlin Heidelberg, 2005.
[3]. Yao, Hong, and Howard J. Hamilton. "Mining functional dependencies from data." *Data Mining and Knowledge Discovery* 16.2 (2008): 197-219.
[4]. Chavan, Anupama A., and Vijay Kumar Verma. "Mining Functional Dependency in Relational Databases using FUN and Dep-Miner: A Comparative Study." *International Journal of Computer Applications* 78.15 (2013).
[5]. K. V. S. V. N. Raju and A. K. Majumdar, "Fuzzy functional dependencies and lossless join decomposition of fuzzy relational database systems," ACM Trans. Database Syst., vol. 13, no. 2, pp. 129–166, 1988.
[6]. M. Arenas and L. Libkin, "A normal form for XML documents," ACM Trans. Database Syst., vol. 29, no. 1, pp. 195–232, 2004.
[7]. M.-L. Lee, T. W. Ling, and W. L. Low, "Designing functional dependencies for XML," in Proc. 8th Int. Conf. Extending Database Technol., 2002, pp. 124–141.
[8]. S.-K. Chang, V. Deufemia, G. Polese, and M. Vacca, "A normalization framework for multimedia databases," IEEE Trans. Knowl. Data Eng., vol. 19, no. 12, pp. 1666–1679, Dec. 2007.
[9]. V. Vianu, "Dynamic functional dependencies and database aging," J. ACM, vol. 34, no. 1, pp. 28–59, 1987.
[10]. U. Nambiar and S. Kambhampati, "Mining approximate functional dependencies and concept similarities to answer imprecise queries," in Proc. 7th Int. Workshop Web Databases, 2004, pp. 73–78.
[11]. D. A. Simovici, D. Cristofor, and L. Cristofor, "Impurity measures in databases," Acta Informatica, vol. 38, no. 5, pp. 307–324, 2002.
[12]. J. Grant and J. Minker, "Normalization and axiomatization for numerical dependencies," Inf. Control, vol. 65, no. 1, pp. 1–17, 1985.
[13]. S. Russell, The Use of Knowledge in Analogy and Induction, ser. Research notes in artificial intelligence. New York, NY, USA: Pitman, 1989.
[14]. M. J. Maher, "Constrained dependencies," Theor. Comput. Sci., vol. 173, no. 1, pp. 113–149, 1997.
[15]. G. Cormode, L. Golab, K. Flip, A. McGregor, D. Srivastava, and X. Zhang, "Estimating the confidence of conditional functional dependencies," in Proc. ACM SIGMOD Int. Conf. Manage. Data, 2009, pp. 469–482.