Mining Approximate Functional Dependencies from Large Databases Based on Concept Similarities to Answer Imprecise Queries

Jismol Abraham and R. Priya

1Vels University, Chennai, India.
jismol2007@gmail.com

2Department of MCA, Vels University, Chennai, India, 
Priyaa.research@gmail.com

Abstract

Nowadays, the information technology has grown rapidly with the evolution of larger databases. Viewing or retrieving information from larger databases seems to be a challenging task. Here comes the role of Functional Dependancy in databases. Functional dependence is a relationship that exists when one attribute uniquely determines another attribute. The discovery of functional dependencies in a dataset has greater importance for database redesign, anomaly detection and data cleansing applications. In this paper, a prototype model/algorithm called SFD (Similarity Functional Dependency) is designed for identifying all the functional dependency values in a dataset for accurate and efficient result verification. This paper elaborates the concept about fetching of data from the database in which there is repeated information or redundancy i.e. database with similar functional values. This concept has been implemented with the help of the similarity functional dependency algorithm, which helps in finding out the redundancy in the database.

Key Words: Data mining, hierarchical algorithm, functional dependency, SFD algorithm, anomaly detection, differential dependencies.
1. Introduction

A database is simply an organized collection of related data, typically stored on disk, and accessible by possibly many concurrent users. The database plays a major role in information technology for storing multi number of values. Due to the increase in the usage level of the database, it is a challenging work to store, view or retrieve values. Among these, retrieving desired files from the large data set is difficult one, as large dataset holds the collection of multi records of related data. In such conditions, we face the challenge for the extraction of hidden predictive information from large databases. This [1. P. A. Flach and I. Savnik] concept describes how to extract the matching or repeated or identical files from the database and to filter them according to their functional tasks. Hospital database is used for filtering the patient details according to their name having identical values.

Data Mining

In simple words, data mining is defined as a process used to extract usable data from a larger set of any raw data. As described in [2. W. Fan, H. Gao, X. Jia, J. Li, and S. Ma] data mining plays a vital role to help companies focus on the most important information in their data warehouses. It mainly involves four classes of tasks such as,

Classification-Classification is one of the data mining technique that assigns categories to a predefined group in order to aid in more accurate predictions and analysis.

Clustering- Clustering is a process of partitioning set of data (or objects) into a set of meaningful sub-classes, called clusters. Clustering help users understand the natural grouping or structure in a data set in grouping the similar number of items.

Regression- Regression is a data mining technique used to find a function which models the data with the least error in the dataset. Regression is used across multiple industries for business and marketing planning, financial forecasting, environmental modeling and analysis of trends.

Association rule learning-Association rule mining searches the relationship between variables, which is also known as a collection of techniques for discovering the efficient unknown, novel, valid patterns in the larger database systems.

Figure1 depicts, how the end-user wants to view or retrieve any data from the database. For retrieving the larger dataset value, the user generates an input value to access that file. In the next step the user input value pattern is evaluated for the purpose of preprocessing.
2. Literature Survey

Existing studies in data mining, mostly focus on finding patterns in large data sets which aids in organizational decision making. This section explains about the papers related to functional dependencies in relational database.

discovery of similarity constraints for matching dependencies, elaborating the application of matching dependencies of various data quality applications such as detecting the violations of integrity constraints. [10]. Debajit Sensarma and Samr Sen Sarma (2015) took a survey on different graph based anomaly detection techniques, providing general, comprehensive, and structured overview of the methods for anomaly detection in data represented through graphs [11]. L.Bertossi, S. Kolahi, and L. Lakshmanan (2011) Data cleaning and query answering with matching dependencies and matching functions. [12].

3. Anomaly Detection

Anomaly detection is the detection or identification of items, events or observations which can’t conform to an expected pattern or other items in a dataset. Normally the anomalous items will translate into some kind of problem such as bank fraud a structural defect, medical problems or errors in a text. Anomalies are also defined as outliers, deviations and exceptions as mentioned in [3. G. Cormode, L. Golab, K. Flip, A. McGregor, D. Srivastava, and X. Zhang]. Figure 3 describes the architecture of anomaly detection. Anomaly detection is done using prediction model of data, after that, if anomalies exists, it will perform clustering and filtering of data.

Unsupervised anomaly detection techniques detect the anomalies in an unlabeled data set under such assumption that the majority of the instances in the data set is normal. Supervised anomaly detection techniques require a data set that has been labeled as normal and abnormal and involves training a classifier. Semi-supervised anomaly detection techniques constructs a model representing the normal behavior of the given training data set and testing the test instance to be generated by the learnt model.

Application of Anomaly Detection

Anomaly detection is used in a variety of domains, such as intrusion detection, fraud detection, fault detection, system health monitoring, event detection in sensor networks, and detecting Ecosystem disturbances. It is frequently used in preprocessing to remove anomalous data from the data set. In supervised
learning [4. I. F. Ilyas, V. Markl, P. Haas, P. Brown, and A. Aboulnaga], removing the anomalous data from the dataset often results in a statistically significant increase in accuracy.

Figure 3: Data Stream Estimator Search

Figure 3 describes that if a user registers his/her information the data stream complexity estimator searches the entire database file. If it finds the same value of already registered user and currently registering user then it fetches the corresponding column values and preprocesses it and then it matches the each and every column values of matched file.

**Differential Dependencies**

Differential Dependencies show the constraints of difference, named as a different function, instead of identification of such types of functions in traditional dependency notations like functional dependencies. Informally, a differential dependency states that if two tuples have distances on attributes X agreeing with the several differential function, then their distances on the attributes should also agree with the same corresponding differential functions on Y as mentioned in [5. Ronald S King and James J. Legendre]. Such type of differential dependencies is useful in various applications such as violation detection, data partition, query optimization and record linkage and so on. Data dependencies, such as Functional Dependencies are explained as in [6.T.Pepenbrock and team.], which are traditionally used for schema design, integrity constraints, and query optimization and so on, with respect to schema quality in databases.[7.D. MurugaRadha Devi and P. Thambidurai] defines Conventional dependencies, originally proposed for schema-oriented issues, which are defined based on equality function, i.e., attribute values are compared according to equality.

4. **Research Methods**

**Data Set**

The set of items/data which are taken into account for this work is hospital database. The dataset is normally defined as the collection of data from a larger amount of database. Data can be fetched as required. A hospital database may
contain a large number of datasets. In those datasets there may be many number of duplicate records or redundant values as described in [8. X. Chu, I. F. Ilyas, and P. Papotti.]. These repeated and duplicate values can be filtered using a similarity dependency algorithm.

**Hierarchical Algorithm**

The hierarchical algorithm is an efficient and normal algorithm. This algorithm [9. S. Kwashie and team] work like a process of FCFS (First Come First Serve). The data which are entered first will be added to the database and it will be stored one by one order or in the sequential order. This method [10. S. Song and L. Chen] has been used frequently in the process of adding data or information to the database easily. In a hierarchal method, separate clusters are finally combined into a single cluster. The main advantage is, it consumes less computation costs. Generally, this algorithm can be divided into two categories, namely divisive and agglomerative. Agglomerative clustering performs the bottom-up strategy, in which it initially considers each and every data point as the singleton cluster. After that, it continues by merging all those clusters into a single cluster. Another method of clustering referred as divisive clustering is used to determine the best approach on how to split the selected clusters into two different new clusters. This method [11. Debajit Sen Sarma and Samr Sen Sarma] creates a hierarchical decomposition of the given set of data objects. Further classification is done through hierarchical methods on the basis of the hierarchical decomposition.

**Similarity Functional Dependency Algorithm**

Similarity functional dependencies (SFD) become a more complex solution for determining the similar datasets in a larger database. The similarity functional dependencies [12. W. Fan and team] defines the statement for finding the data or files with the same number of repeated values in databases. For example, consider the hospital database, in which there may be a large number of records in the database. Retrieving the details of one particular patient is quite difficult because the entire database may contain many numbers of patients having the same name. Not only that [15. A. K. Elmagarmid, P. G. Ipeirotis, and V. S. Verykios] describes that the name is same but there may be a chance of having a patient with same name, same date of birth, same blood group and also a chance of having the same admission date to those corresponding hospitals. So sometimes there might be a manual mistake in providing treatments for those different patients having same records. In such conditions, the similarity functional dependency algorithm is useful for fetching those similar records and separate them correspondingly. [16. J. Liu, J. Li, C. Liu, and Y. Chen] surveys the challenges and also finds the possible solutions to these problems. If the structure of the information to be searched is sufficiently simple, for instance in single-dimensional numerical attributes or character strings, such problems can be considered as easy to solve. Recently, an increasing number of applications [18. M. _I. Seozat and A. Yazici] have emerged for the purpose of processing the enormous amounts of application specific data objects.
5. Proposed Method

Algorithm for Similarity Functional Dependencies (SFD)

INPUT:
Hospitals dataset is the input value, minimum support factor, confidence factor.

OUTPUT:
Frequent item sets which predicts repeated and duplicate records in the dataset.

METHOD:
Dtrue: X1; X2X1; X2
Cerr0p ____Y1; Y2Y1; Y2;
Where uX1; X2
UAi; Aj and uY1; Y2
UBk; Bl,
with UAi 2 X1, Aj 2 X2, Bk 2 Y1, Bl 2 Y2; UAi;Aj
Where Bl is the comparison function in the data space S.
UAi- minimum support factor
Minimum support factor is a form of database intervention, which is having the
hierarchical input values used for preprocessing.
UBk-confidence factor
Confidence factor having preprocessed and separated values
Aj- used just for storing the preprocessed values of minimum support factor
UAi.

uX1—used as a variable for storing the values of X1, X2.
uY1—handles the separated values of Y1 and Y2 in order to arrange the values
into different categories according to their matched parent values.
1. Perform basic pre-processing with the hospital database.
2. Check the pre-processed value Y1 and Y2 value with the input values.
3. Calculate number of records matching with the input values X1 and X2.
4. After that fetch the similar records from the database having same X1, X2,
   Y1, Y2 values.
5. Finally the segmentation is done into different group of similar records for
   further operations.
The result is depicted as given below:

<table>
<thead>
<tr>
<th>Input</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 4: Input Value</td>
<td>Figure 5. Result</td>
</tr>
</tbody>
</table>
Results and Discussion

In this research work, an attempt is done to discover the similarities in the repeated and duplicate dataset values of particular data from the hospital database. The segmentation of similar records is done by designing a new algorithm called Similarity Functional Dependancy. As per the new algorithm, the data comparison is made quite easy as a new record is being compared with the preprocessed segmented data which is created.

This preprocessed segmentation can outperform the existing algorithms related to functional dependency. Anomaly detection methodology is implemented for finding repeated value from the huge database.

Future Work

Future work is to fetch the repeated and similarity functional values dynamically using other techniques of data mining such as regression, clustering and neural network in order to find similarities among multiple attributes.

References


