Chapter 9

Mining Rare Association Rules by Discovering Quasi-Functional Dependencies: An Incremental Approach

Giulia Bruno
Politecnico di Torino, Italy

Paolo Garza
Politecnico di Torino, Italy

Elisa Quintarelli
Politecnico di Milano, Italy

ABSTRACT

In the context of anomaly detection, the data mining technique of extracting association rules can be used to identify rare rules which represent infrequent situations. A method to detect rare rules is to first infer the normal behavior of objects in the form of quasi-functional dependencies (i.e. functional dependencies that frequently hold), and then analyzing rare violations with respect to them. The quasi-functional dependencies are usually inferred from the current instance of a database. However, in several applications, the database is not static, but new data are added or deleted continuously. Thus, the anomalies have to be updated because they change over time. In this chapter, we propose an incremental algorithm to efficiently maintain up-to-date rules (i.e., functional and quasi-functional dependencies). The impact of the cardinality of the data set and the number of new tuples on the execution time is evaluated through a set of experiments on synthetic and real databases, whose results are here reported.

INTRODUCTION

Mining rare events is a critical task in several research areas such as database, machine learning, knowledge discovery, fraud detection, activity monitoring, and image analysis. In all these situations the main goal is to identify objects of a given population whose behavior is anomalous with respect to a set of rules part of the knowledge base, usually expressed by means of some statistical kind of computation on the given population. These exceptional situations
are usually referred as outliers in the literature. Indeed, “an outlier is an observation that lies an abnormal distance from other values in a random sample from a population” ("NIST/SEMATECH" 2007). In this work, we define such infrequent situations as anomalies.

Anomalies indicate abnormal conditions, such as anomalous objects in an image, intruders inside a system, and faults on a factory production line (Hodge & Austin, 2004). It is imperative to detect sudden changes in the patterns which may indicate problems or exceptions in the behaviours. Anomaly detection methods can monitor individual shares or markets and detect novel trends which may indicate buying or selling opportunities. In a database, anomalies may indicate fraudulent cases or they may denote an error by the entry clerk or a misinterpretation of a missing value code; apart from the method, the detection of the anomaly is vital for database consistency and integrity.

The problem of analyzing anomalies is interesting, since they represent errors or semantically correct, albeit infrequent, situations. In both cases, detecting anomalies is a challenging task either to allow one to correct erroneous data or to investigate the meaning of exceptions. In the database research field and in real applications, the problem of identifying and correcting anomalies has received a lot of attention in recent years (Angiulli et al., 2007; Angiulli & Pizzuti, 2005; Ceri et al., 2007). The outlier detection problem is considered in order to examine a target database to derive more complex forms of constraints, with respect to those that are imposed during the design phase. This task is of great importance also in a variety of application fields, especially for biological, financial, and clinical data, where it is important to detect anomalies in order to clean errors or discover exceptions needing a further investigation by specialists (Apiletti et al., 2006).

A method to detect rare rules, which represent anomalies, is to first infer the normal behavior of objects by extracting frequent rules from a given dataset, and then analyzing rare violations to what has been inferred as a normal behavior (Bruno et al., 2007). The frequent rules are described in the form of quasi-functional dependencies and mined from the dataset by using association rules. A quasi-functional dependency is an approximate functional dependency derived from data (Baralis et al., 2004), and represents an implication among attributes (in the context of relational databases) or among elements (with respect to XML documents), which frequently holds in the analyzed dataset. Each quasi-functional dependency is characterized by a dependency degree value that represents the strength of the dependency. For each mined quasi-functional dependency, the set of association rules that involve the same attributes is queried in order to detect anomalies. In particular, the rare association rules (i.e., with a confidence lower than a fixed threshold) are selected, and further investigated in order to distinguish interesting anomalies from erroneous data.

Quasi-functional dependencies, which are a priori unknown, are usually inferred from the current instance of the considered database. However, in several applications anomalies have to be promptly detected, but the instance continuously changes (i.e., augment) over time (e.g., financial data). A more recent scenario where dynamicity is fundamental is the one where high level information processing tasks, like environmental monitoring, tracking, and classification, are performed with wireless sensor networks, where sensors must process a continuous (possibly fast) stream of data (Mozafari et al., 2008). When monitoring time critical activities, outlier detection problems have to be considered, in order to isolate erroneous or significant, but infrequent, streams of data. In this dynamic scenario all the inferred quasi-functional and functional dependencies, their dependency degree value, and the set of anomalies can potentially change over time.