Sebastián Ventura · José María Luna

Pattern Mining with Evolutionary Algorithms

Springer

Pattern Mining with Evolutionary Algorithms

Sebastián Ventura • José María Luna

Pattern Mining with Evolutionary Algorithms

Springer

Sebastián Ventura
Department of Computer Science
and Numerical Analysis
University of Cordoba
Cordoba, Spain

José María Luna
Department of Computer Science
and Numerical Analysis
University of Cordoba
Cordoba, Spain

ISBN 978-3-319-33857-6 ISBN 978-3-319-33858-3 (eBook)
DOI 10.1007/978-3-319-33858-3

Library of Congress Control Number: 2016939025

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG Switzerland

–Success is a journey, not a destination. The doing is often more important than the outcome–

Arthur Ashe.

To our families.

Preface

This book is intended to provide a general and comprehensible overview of the field of pattern mining with evolutionary algorithms. To do so, the book provides formal definitions about patterns, pattern mining, type of patterns, and the usefulness of patterns in the knowledge discovery process. As it is described within the book, the discovery process suffers from both high runtime and memory requirements, especially when high-dimensional datasets are analyzed. To solve this issue, many pruning strategies have been developed. Nevertheless, with the growing interest in the storage of information, more and more datasets comprise such a dimensionality that the discovery of interesting patterns becomes a hard process. In this regard, the use of evolutionary algorithms for mining pattern enables the computation capacity to be reduced, providing sufficiently good solutions.

The book also provides a survey on evolutionary computation with particular emphasis on genetic algorithms and genetic programming. Additionally, this book carries out an analysis of the set of quality measures most widely used in the field of pattern mining with evolutionary algorithms. This book serves as a good review on the most important evolutionary algorithms for pattern mining. In this sense, it considers the analysis of different algorithms for mining different types of patterns and relationships between patterns, such as frequent patterns, infrequent patterns, patterns defined in a continuous domain, or even positive and negative patterns.

The book also introduces a completely new problem in the pattern mining field, which is known by the name of the mining of exceptional relationships between patterns. In this problem, the goal is to identify patterns where distribution is exceptionally different from the distribution in the complete set of data records. Finally, this book deals with the subgroup discovery task, a method to identify a subgroup of interesting patterns that is related to a dependent variable or target attribute. This subgroup of patterns satisfies two essential conditions: interpretability and interestingness.

Cordoba, Spain
February 2016

Sebastián Ventura
José Marí a Luna

Acknowledgments

We would like to thank the Springer editorial team for giving us the opportunity to publish this book, for their great support toward the preparation and completion of this work, and for their valuable editing suggestions to improve the organization and readability of the manuscript. We also want to thank our colleagues for their valuable help during the preparation of the book, whose comments were very helpful for improving its quality.

This work was supported by the Spanish Ministry of Economy and Competitiveness under the project TIN2014-55252-P and FEDER funds.

Contents

Chapter 1
Introduction to Pattern Mining

Abstract This chapter introduces the pattern mining task to the reader, providing formal definitions about patterns, the pattern mining task and the usefulness of patterns in the knowledge discovery process. The utility of the extraction of patterns is introduced by a sample dataset for the market basket analysis. Different type of patterns can be considered from the pattern mining point of view, so an exhaustive taxonomy about patterns in this field is presented, describing concepts such as frequent and infrequent patterns, positive and negative patterns, patterns expressed in compressed forms, sequential patterns, spatio-temporal patterns, etc. Additionally, some pruning strategies to reduce the computational complexity are described, as well as some efficient pattern mining algorithms. Finally, this chapter formally describes how interesting patterns can be associated to analyse the causality by means of association rules.

1.1 Definitions

The exponential increasing amounts of data generated and stored in many different areas brings about the need for analysing and getting useful information from that data. Generally, raw data lacks of interest and an in-depth analysis is required to extract the useful knowledge that is usually hidden. This process has giving rise to the field known as knowledge discovery in databases (KDD) [12].

KDD is concerned with the development of methods and techniques for making sense of data, and the discovery of patterns plays an important role here. Under the term *pattern* we can define subsequences, substructures or itemsets that represent any type of homogeneity and regularity in data [1]. Thus, patterns represent intrinsic and important properties of datasets.

Given a set of items $I = \{i_1, i_2, \ldots, i_n\}$ in a dataset, a pattern P is formally defined as a subset of I, i.e. $\{P = \{i_j, \ldots, i_k\} \subseteq I, 1 \leq j, k \leq n\}$, that describes valuable features of data. Given a pattern P, the length or size of that pattern is expressed as $|P|$, denoting the number of single items or singletons that it includes. Thus, the length of a simple pattern $P = \{i_1, \ldots, i_j\} \subseteq I$ is defined as $|P| = j$ since it is comprised of j singletons.

S. Ventura, J.M. Luna, *Pattern Mining with Evolutionary Algorithms*,
DOI 10.1007/978-3-319-33858-3_1

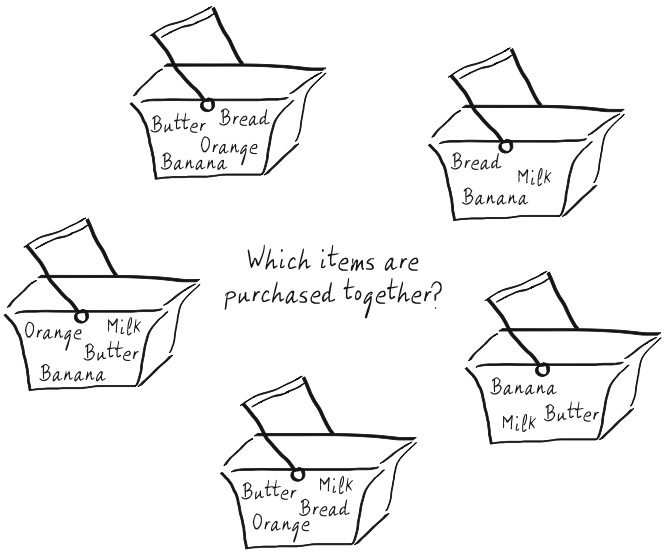

Fig. 1.1 Analysis of the market basket, including five different customers

The task of finding and analysing patterns in a database might be considered as straightforward [2], but it gets increasingly complicated once the interest of the discovered patterns becomes a priority. Pattern mining is considered as an essential part of the KDD process, being defined as the task of discovering patterns of high interest in data for a specific user aim. Thus, the process of quantifying in a proper way the interest of the discovered patterns includes different metrics that are highly related to the purpose for which the task is applied [9]. Market basket analysis is perhaps the first application domain in which mining patterns of interest was useful. In many occasions, a high number of purchases are bought on impulse [6] and it is required an in-depth analysis to obtain clues that let us know which specific items are strongly related. For instance, how likely is for a customer to buy bananas and milk together? In this regard, it seems easy to determine that the interest of the patterns might be quantified by their probability of occurrence [12]. Thus, considering a sample set of customers (see Fig. 1.1), we can obtain that the pattern $P = \{Banana, Milk\}$ has a frequency of three over a total of five customers, so there exists a probability of 60 % that a random customer buys both bananas and milk in the same trip. This analysis could allow shopkeepers to increase the sales by re-locating the products on the shelf, or even it could allow managers to plan advertising strategies.

Pattern mining results can be analysed in many different ways, and the same pattern might be used to follow different marketing strategies. For instance, the aforementioned pattern $P = \{Banana, Milk\}$ might be used to understand that placing these products in proximity produces an encouragement of the sales of both products at time. Any customer that buys any of these items will have the

temptation of buying the other one. Nevertheless, the same pattern P might be analysed differently, and both products might be placed far from each other. In this regard, since both products tend to be purchased together, the fact of a customer going over the supermarket may give rise to buy something else on impulse.

According to some authors [1], pattern mining is a high level and challenging task due to the computational challenge issues. The pattern mining problem turns into an arduous task depending on the search space, which exponentially grows with the number of single items or singletons considered into the application domain. For a better understanding, let us consider a set of items $I = \{i_1, i_2, \ldots, i_n\}$ in a dataset. A maximum number of $2^{|I|} - 1$ different patterns can be found, so a straightforward approach becomes extremely complex with the increasing number of singletons. As a matter of example, let us consider that, in average, about 8,000 grains of sand fit in a cubic centimetre. Hence, it is estimated that there are about 10×10^{36} grains of sand on all the world's beaches. This value is lower than the number of patterns that can be found in a dataset comprising 150 singletons, i.e. $2^{150} - 1 \approx 1.42 \times 10^{45} \gg 10 \times 10^{36}$.

1.2 Type of Patterns

Since its introduction [5], the pattern mining task has provoked tremendous curiosity among experts on different application fields [26, 28]. This growing interest encouraged the definition of new type of patterns according to the analysis required by experts on different application field. The following discussion will address and describe each of these patterns.

1.2.1 Frequent and Infrequent Patterns

Market basket analysis is considered as the first application domain in which the pattern mining task plays an important role. In this domain, the mining of strongly related products is essential to take right decisions to increase the sales, so the mining of frequent products purchased together is the key [3]. To quantify this frequency, it is calculated as the number of transactions in which the set of products is included in a dataset. Table 1.1 illustrates an example dataset based on five different customers (see Fig. 1.1). Each row corresponds to a different transaction that represents a customer, and each transaction contains a set of items that describes the products purchased by the customer.

Let $I = \{i_1, i_2, \ldots, i_n\}$ be a set of items in a dataset, and let $T = \{t_1, t_2, \ldots, t_m\}$ be the set of all dataset transactions. Each transaction t_j is a set of items such that $t_j \subseteq I$. Let P also be a pattern comprising a set of items, i.e. $P \subseteq I$. The pattern P satisfies the transaction t_j if and only if $P \subseteq t_j$, and the frequency of this pattern $f(P)$ is defined as the number of different transactions that it satisfies, i.e.

Table 1.1 An example of market basket dataset obtained from Fig. 1.1

Customer	Items
ID_1	{Banana, Bread, Butter, Orange}
ID_2	{Banana, Bread, Milk}
ID_3	{Banana, Butter, Milk}
ID_4	{Bread, Butter, Milk, Orange}
ID_5	{Banana, Butter, Milk, Orange}

$f(P) = |\{\forall t_j \in T : P \subseteq t_j\}|$. A pattern P is defined as frequent if and only if the number of transactions that it satisfies is greater or equal to a minimum predefined threshold f_{min}, i.e. $f(P) \geq f_{min}$.

As previously analysed, a dataset comprising n singletons contains $2^n - 1$ different itemsets, whereas the number of itemsets of size k is equal to $\binom{n}{k}$ for any $k \leq n$. Thus, given the amount of computations needed for each candidate is $O(k)$, the overall complexity of the mining is $O(\sum_{k=1}^{n} k \times \binom{n}{k}) = O(2^{n-1} \times n)$. All of this led us to the conclusion that the complexity of finding patterns of interest is in exponential order, and this complexity is even higher when the frequency of each itemset is calculated [14]. For any dataset comprising n singletons and m different transactions, the complexity to compute the frequencies of all the patterns within the dataset is equal to $O(2^{n-1} \times m \times n)$.

Going back to the aforementioned example, a space of exploration of size $2^5 - 1 = 31$ can be arranged as a lattice (see Fig. 1.2) where the number of transactions satisfied by each itemset is illustrated into brackets. The number of k-itemsets (itemsets of size k) that can be found in a dataset that comprises n singletons is $\binom{n}{k} = \frac{n!}{k!(n-k)!}$. Thus, considering the sample market basket dataset, the number of 1-itemsets is $\binom{5}{1} = 5$; the number of 2-itemsets is $\binom{5}{2} = 10$; the number of 3-itemsets is $\binom{5}{3} = 10$; etc, obtaining that $\binom{5}{1} + \binom{5}{2} + \ldots + \binom{5}{5} = 2^5 - 1 = 31$, which is the whole space of exploration for this example.

Continuing with the same example, and considering a $f_{min} = 2$, the lattice is separated (dashed line) into two main parts: frequent and infrequent. Any pattern P that appears above the border that divides the search space into two parts is considered as a frequent pattern since $f(P) \geq f_{min} = 2$. Thus, according to this measure, which determines whether a pattern is of interest or not, the space of exploration can be reduced significantly. For instance, from a total of $2^5 - 1 = 31$ itemsets, the search space is reduced to a total of 19 frequent patterns when $f_{min} = 2$ is considered. The increasing of f_{min} implies a reduction of the search space.

As mentioned above, the process of mining frequent patterns obtains a set of patterns whose frequency overcomes the minimum threshold f_{min}. According to f_{min}, any frequent pattern represents the most valuable information that reflect intrinsic and important properties of datasets. However, there are situations where it is interesting to discover abnormal or unusual behaviour in datasets, discovering rare or infrequent patterns [15], i.e. those that do not follow the trend of the others [33]. A pattern P is defined as infrequent if and only if the number of transactions that

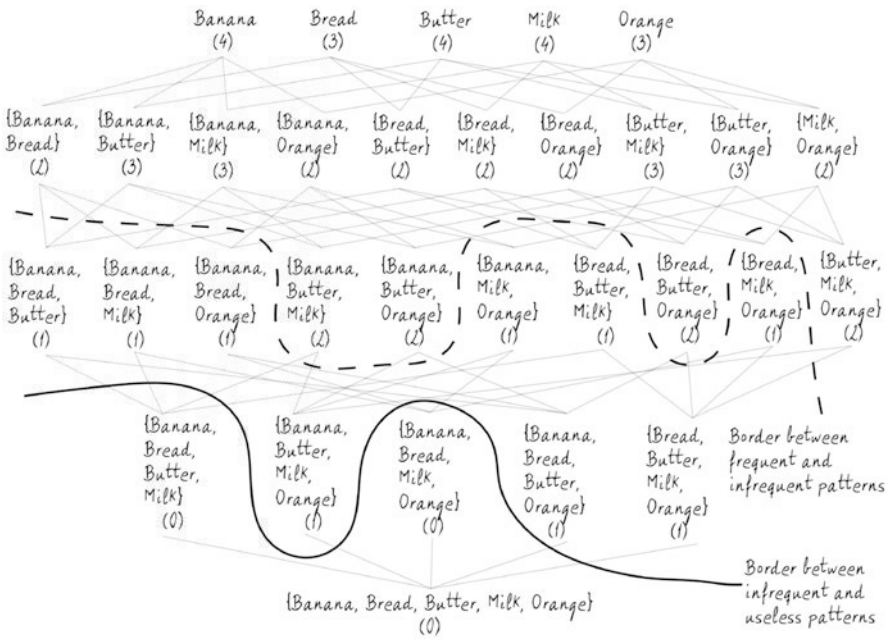

Fig. 1.2 Lattice analysis of the market basket, including five different customers

it satisfies is lower than a maximum predefined threshold f_{max}, i.e. $f(P) = |\{\forall t_j \in T : P \subseteq t_j\}| < f_{max}$. Nevertheless, this maximum threshold value implies that any itemset that does not appear in the dataset will also be considered as infrequent. This can be solved by including a minimum threshold that divides the set of non-frequent patterns into infrequent and useless patterns—those that can be discarded. Thus, a second definition of infrequent patterns can be described. A pattern P is defined as infrequent if and only if the number of transactions that it satisfies is lower than a maximum predefined threshold f_{max} and greater than a minimum threshold f_{min}, i.e. $f_{min} < f(P) < f_{max}$.

Again analysing the lattice (see Fig. 1.2) that comprises the itemsets of the sample market basket dataset illustrated in Table 1.1, and considering $f_{max} = 2$, and $f_{min} = 0$, it is obtained that the set of infrequent patterns comprises those patterns located between the dashed (border between frequent and infrequent patterns) and the solid (border between infrequent and useless patterns) lines. Thus, according to both thresholds, there are nine infrequent patterns in the dataset. Finally, we obtain three patterns that are not included in the dataset and they can be considered as useless.

1.2.2 Closed and Maximal Frequent Patterns

Many real-world application domains [22] contain patterns whose length is typically too high. The extraction of such lengthy patterns requires a high computational time when the aim is the extraction of frequent patterns. Noted that a major feature of any frequent pattern is that any of its subsets is also frequent [1], so a significant amount of time can be spent on counting redundant patterns. For instance, given a pattern P of length $|P| = 5$, it comprises a total of $2^5 - 2 = 30$ sub-patterns that are also frequent. In this regard, the discovery of condensed representations of frequent patterns is a solution to overcome the computational and storage problems.

Suppose a dataset whose set of frequent patterns is defined as $\mathscr{P} = \{P_1, P_2, \ldots, P_n\}$. A frequent pattern $P_i \in \mathscr{P}$ is defined as maximal frequent pattern if and only if it has no frequent superset, i.e. $\{P_i : \nexists P_j \supset P_i, P_j \in \mathscr{P} \wedge P_i \in \mathscr{P}\}$. The number of maximal frequent patterns is considerably smaller than the number of all frequent patterns. For the sample market basket dataset shown in Table 1.1 the set of maximal frequent patterns is illustrated in Fig. 1.3. As illustrated, there is no super-patterns for the pattern $P = \{Banana, Bread\}$ that satisfies the minimum frequency threshold $f_{min} = 2$, so P is defined as a maximal frequent pattern.

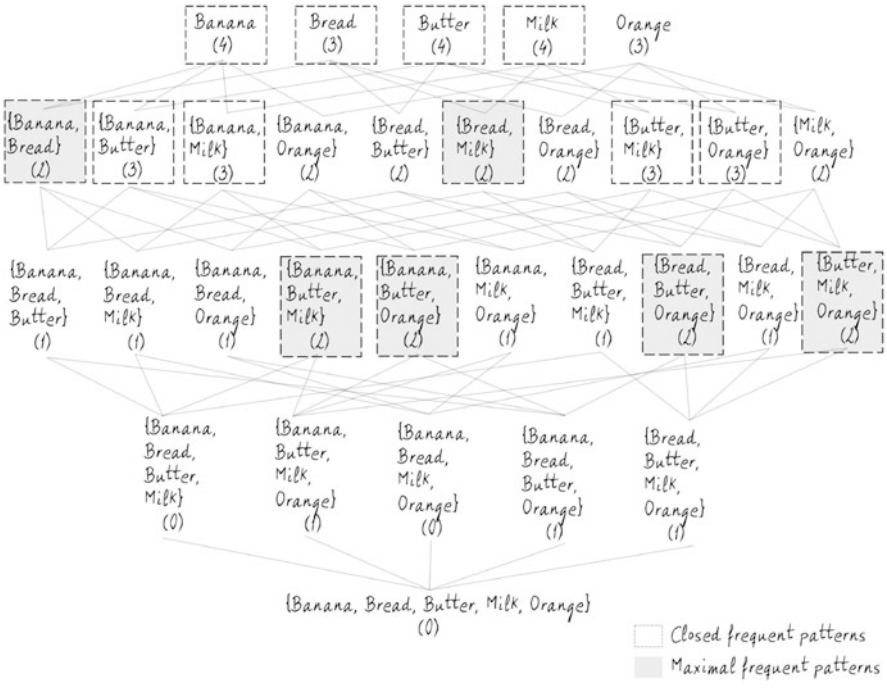

Fig. 1.3 Maximal and closed frequent patterns represented in the lattice of the market basket dataset