Visualization of Numerical Association Rules by Hill Slopes

Iztok Fister Jr.[1], Dušan Fister[2] Andres Iglesias[3,4], Akemi Galvez[3,4], Eneko Osaba[6], Javier Del Ser[5,6], and Iztok Fister[1,3]

[1] Faculty of Electrical Engineering and Computer Science, University of Maribor, Smetanova 17, 2000 Maribor, Slovenia iztok.fister1@um.si
[2] Faculty of Economics and Business, University of Maribor, Razlagova 14, 2000 Maribor
[3] University of Cantabria, Avenida de los Castros, s/n, 39005 Santander, Spain
[4] Toho University, 2-2-1 Miyama, 274-8510, Funabashi, Japan
[5] University of the Basque Country (UPV/EHU), Bilbao, Spain
[6] TECNALIA, Basque Research and Technology Alliance (BRTA), Derio, Spain.

Abstract. Association Rule Mining belongs to one of the more prominent methods in Data Mining, where relations are looked for among features in a transaction database. Normally, algorithms for Association Rule Mining mine a lot of association rules, from which it is hard to extract knowledge. This paper proposes a new visualization method capable of extracting information hidden in a collection of association rules using numerical attributes, and presenting them in the form inspired by prominent cycling races (i.e., the Tour de France). Similar as in the Tour de France cycling race, where the hill climbers have more chances to win the race when the race contains more hills to overcome, the virtual hill slopes, reflecting a probability of one attribute to be more interesting than the other, help a user to understand the relationships among attributes in a selected association rule. The visualization method was tested on data obtained during the sports training sessions of a professional athlete that were processed by the algorithms for Association Rule Mining using numerical attributes.

Keywords: Association Rule mining · Optimization · Sports training · Tour de France · Visualization

Citation details: https://doi.org/10.1007/978-3-030-62362-3_10

1 Introduction

Association Rule Mining (ARM) [1] is an important part of Machine Learning that searches for relations among features in a transaction database. The majority of the algorithms for ARM work on categorical features, like Apriori proposed by Agrawal [2], Eclat, introduced by Zaki et al. [20], and FP-Growth, developed by Han et al. [8]. Algorithms for dealing with the numerical features have also been developed recently [3, 6].

Typically, these algorithms generate a huge number of association rules, from which it is hard to discover the most important relations. In order to extract a knowledge from the collection of mined association rules, a lot of tools have emerged [7] that are able to cope with complex data structures, e.g., instance-relationship data [5], user generated context [9, 12], and scanner data [14]. On the other hand, there are a number of papers proposing data visualization methods as a means for extracting meaningful results from highly complex settings [4].

This paper focuses on the visualization of ARM using numerical features. Thus, the relationships among these features are visualized using an inspiration taken from one of the most prominent cycling races in the world, i.e., the Tour de France (TDF). Similar as in the Tour de France cycling race, where the hill climbers have more chances to win the race, when the race contains more hills to overcome, the virtual hill slopes, reflecting a probability of one attribute to be more interesting than the other, help a user to understand the relationships among attributes in a selected association rule.

The proposed visualization method was applied on a transaction database consisting of data obtained by measuring data during the training sessions with a mobile device worn by a professional cyclist in the past three seasons. The results of visualization reveal that using the method in the real-world can improve the interpretation of the mining of the association rules, and direct the user to the more important ones.

The structure of the paper is as follows. Section 2 highlights the basic information needed for understanding the subject of the paper. In Section 3, an algorithm for visualizing the mined association rules is illustrated in detail. The results of the proposed visualization method are presented in Section 4. The paper is concluded by Section 5, which summarizes the work performed and outlines directions for the future.

2 Basic information

This section is focused on the background information necessary for understanding the subject that follows. At first, the principles of TDF serving as an inspiration for visualization are highlighted, followed by describing the problem of discovering association rules.

2.1 Tour de France

Numerous research articles published in the past support the idea that good hill climbing abilities are a nuisance for winning the Tour De France (TDF). Climber specialists (fr. grimpeur), all-rounders (fr. rouleur) and time-trial specialists (fr. chronoman) usually fight for overall podium positions at the Champs-lyses, contrary to the breakaway specialists (fr. baroudeur) and sprinters, who strive for glory at individual stages. Good hill climbing abilities come naturally come with suitable anthropometric stature: According to [16], climbers usually weigh 60-66 kg, with their BMI (Body Mass Index) reaching 19-20 kg/m^2. Compared to other specialists, climbers perform exceptionally well at maintaining high relative power output to their weight W/kg. From the characteristics of climbers, we deduce that steep climbs, where intensity is near maximum, determine (or are crucial for) the overall winner of TDF, and we further deduce that climbers are in a favorable role there.

Lucia at al. [10] introduce the TDF historical overview and climbing facts. Good climbing performance in the Alps and Pyrenees is highly correlated to good time-trial performance [18], which provides a good chance that a strong climber will perform solidly at the time-trial, too. Indeed, both are necessary to win the TDF [17]. However, Rogge et al. [13] agree that good performance at mountain stages is crucial for a favorable TDF result overall. Torgler [18] further exposes the difficulties of mountain stages, and emphasizes the high efforts needed to provide good General Classification (GC); the steep mountainous stages are supposed to be the most difficult ones among them all and, thus, are decisive for a successful GC; subsequent climbs with descents are found to be most exhaustive. It follows that, the more exhaustive the stage, the larger the time differences at the finish. Sanders and Heijboer [15] supported this idea, by finding out that mountain stages are of the highest intensity and exercise load among mass-start stage types. To a higher degree, this is because of the total elevation gain and highly alternating pace that occurs at hilltops. In our opinion, van Erp et al. [19] present the most comprehensive empirical study of individual load, intensity and performance characteristics of a single GC contender. Among many testing hypotheses, authors state that the most necessary to compete for victory in a Grand Tour is to achieve the highest power output possible (app. 5.7 - 6.0 W/kg) at key mountain stages.

Climbing to "Hors Catgorie" (HC) climbs is extremely specific. Such climbs are usually of extraordinary distance and elevation, and, thus, require extreme efforts. At high altitudes, moderate hypoxia can come into play, which tightens the cyclist's margins and increases physical fatigue even more. An example of an HC finish climb is shown in Fig. 1. These facts contribute easily to early exhaustion, or overreaching [10] and thus are critical for good overall GC classification. The lost time at mountainous stages usually cannot be recovered anymore (cyclists can barely limit only the losses).

On the other hand, competing to win the TDF is not only about climbing. The TDF is extremely psychologically and physically demanding, especially for GC contenders: (1) These need to be cautious of opponents at all times, (2) No relaxation days are allowed to them and (3) a single bad day or opponent's explosive burst may be devastating. Without mentioning high temperatures, team spirit, injuries, crashes and technical glitches, the psychological and physical tension to GC contenders are the highest at the high-intensity phases, such as steep hills [11].

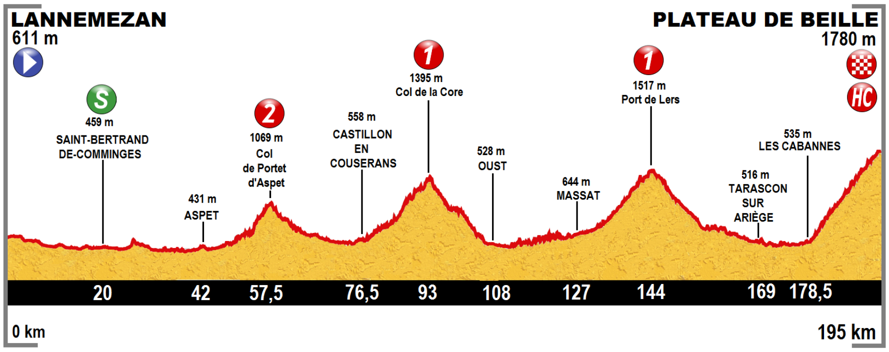

Fig. 1: Example of TDF Stage 12, TDF 2015. Image origin: `https://commons.wikimedia.org/wiki/File:Profile_stage_12_Tour_de_France_2015.png`, distributed under: Creative Commons Attribution-Share Alike 4.0 International License

2.2 Association Rule Mining

ARM can be defined formally as follows: Let us assume a set of objects $O = \{o_1, \ldots, o_M\}$ and transaction dataset $D = \{T\}$ are given, where each transaction T is a subset of objects $T \subseteq O$. Then, an association rule is defined as an implication:

$$X \Rightarrow Y, \tag{1}$$

where $X \subset O$, $Y \subset O$, and $X \cap Y = \emptyset$. In order to estimate the quality of a mined association rule, two measures are defined: A support and a confidence. The support is defined as:

$$supp(X) = \frac{|t \in T; X \subset t|}{|T|}, \tag{2}$$

while the confidence as:

$$conf(X \Rightarrow Y) = \frac{n(X \cup Y)}{n(X)}, \tag{3}$$

where function $n(.)$ calculates the number of repetitions of a particular rule within D, and $N = |T|$ is the total number of transactions in D. Let us emphasize that two additional variables are defined, i.e., the minimum confidence C_{min} and the minimum support S_{min}. These variables denote a threshold value limiting the particular association rule with lower confidence and support from being taken into consideration.

3 Description of constructing the new visualization method for ARM

The problem of predicting the winner of the TDF can be defined informally as follows: Let us assume that cyclist X is a specialist for hill climbing. This cyclist started in n from the total N races and overcame a set of elite cyclists $Y_1, \ldots, Y_{m-1}$ $M_1, \ldots, M_{m-1}$-times, respectively, where m denotes the number of observed cyclists, i.e., not only hill climbers. Such framework could be applied to all cyclists in general. Based on these data, the question is, what is the probability for X to win the TDF this year?

The problem could be solved visually in the sense of the ARM as follows: The number of races in which the cyclist X started can be expressed as $supp(X)$. The same is also true for cyclist Y_i, where the number of his starts is expressed as $supp(Y_i)$. The numbers of races, where X overcame Y_i can be expressed as $conf(X \Rightarrow Y_i)$. Then, the equivalent relation

$$supp(X) \equiv conf(X \Rightarrow Y) \tag{4}$$

means that cyclist X overcame cyclist Y_i in all races in which they both started. This relation can be visualized as a rectangular triangle with two sides of equal length, $supp(Y_i)$ and $conf(X \Rightarrow Y_i)$, and the length of diagonal L is expressed by Pythagoras rule, as follows:

$$L = \sqrt{supp^2(Y_i) + conf^2(X \Rightarrow Y_i)}. \tag{5}$$

As a result, a sequence of triangles is obtained, where each triangle highlights the relationship between the two cyclists. If triangles are ordered according to their supports and put onto a line with spacing proportional to a $conf(X \Rightarrow Y_i)$, and the model triangle with two sides equal to $supp(X)$ is added, the new visualization method emerges, as presented in Fig. 2.

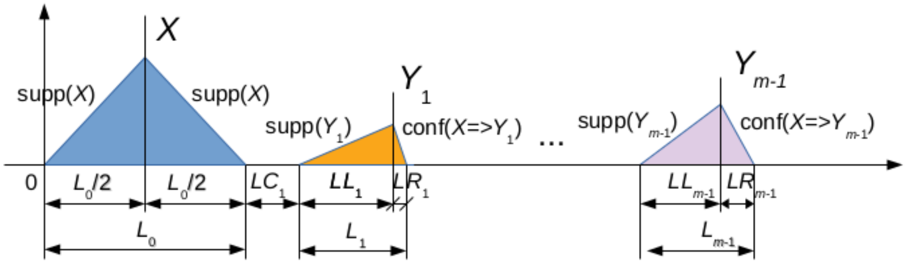

Fig. 2: Mathematical model of virtual hills, on which the new visualization method is founded.

As can be seen from Fig. 2, the model triangle representing the triangle with the largest area, is isosceles, and it is placed on the position $L_1/2$ distant from the origin of 0. However, the positions L_i for $i = 0, \ldots, m-1$ are calculated according to Eq. 5. A corresponding length of the triangle is calculated for each of the observed cyclists Y_i. The position of the triangle on the line is determined as follows:

$$pos_i = L_0 + \sum_{j=1}^{i-1} (LC_j + LL_j + LR_j) + (LC_i + LL_i)\,, \tag{6}$$

where L_0 denotes the diagonal length of the model triangle, $LC_j \propto conf(X \Rightarrow Y_j)$ is the distance between two subsequent triangles, LL_j, expressed as follows:

$$\begin{aligned} \cos\alpha &= \frac{supp(Y_j)}{L_j}, \quad \text{for } j = 1, \ldots, m-1, \\ LL_j &= supp(Y_j) \cdot \cos\alpha = \frac{supp^2(Y_j)}{L_j}, \end{aligned} \tag{7}$$

while LR_j as:

$$\begin{aligned} \cos\beta &= \frac{conf(X \Rightarrow Y_j)}{L_j}, \quad \text{for } j = 1, \ldots, m-1, \\ LR_j &= conf(X \Rightarrow Y_j) \cdot \cos\beta = \frac{conf^2(X \Rightarrow Y_j)}{L_j}. \end{aligned} \tag{8}$$

The interpretation of these triangles representing hills in the TDF is as follows: At first, the larger the area of the triangle for cyclist Y_i, the more chances for cyclist X to overcome this in the race. The same relationship is highlighted in the distance between these triangles. This means that the more the triangle is away from the model triangle, the higher is the probability that X will be overcome by Y_i. Let us emphasize that the discussion is valid for classical ARM using numerical attributes, where the mined association rules serve only as a basis for determining the best features. Thus, the implication relation in the rule is ignored and redefined by introducing the $m-1$ individual implication relations between pairs of features.

Similar as steep slopes have a crucial role in the TDF cycle race for determining the final winner, the virtual hill slopes help the user to understand the relations among features in the transaction database. Indeed, this visualization method can also be applied for visualizing features in the ARM transaction databases. Here, the features are taken into consideration instead of cyclists. On the other hand, interpretation of visualization is also slightly different from the TDF. Here, we are interested in those relations between features that most highlights the mined association rules. The larger the area of the triangle Y_i, the closer the relationship between the feature X.

4 Experiments and results

The goal of our experimental work was to show that the mined association rules can be visualized using the proposed visualization method based on inspiration taken from the TDF cycling competition. In line with this, two selected association rules mined from a corresponding transaction database are visualized.

The association rules were mined using contemporary approaches using stochastic nature-inspired population-based algorithms, like Differential Evolution [6]. Then, the best mined association rules according to support and confidence are taken into consideration. In each specific association rule, the feature with the the best support is searched for. This feature serves as a model, with which all the other features in the observed rule are compared according to a confidence.

The transaction database consists of seven numerical features, whose domain of feasible values are illustrated in Table 1. The transaction database consists of 700 transactions representing the results measured by a mobile device worn by a professional cyclist during the sports training session. These data were obtained in the last three cycling seasons. In a temporary sense, this means that we can start to set a get valuable visualization with full predictive power for the current season after lapse of three seasons. However, this does not mean that the method cannot be applied before elapsing three seasons, but the obtained results could be less accurate.

Table 1: Observed numerical features with their domain of values.

Nr.	Feature	Domain
F-1	Duration	[43.15, 80.683]
F-2	Distance	[0.00, 56.857]
F-3	Average HR	[72, 151]
F-4	Average altitude	[0.2278, 1857.256]
F-5	Maximum altitude	[0.0, 0.0]
F-6	Calories	[20.0, 1209]
F-7	Ascent	[0.00, 1541
F-8	Descent	[0.00, 1597]

The best two association rules according to support are presented in Table 2, where they are denoted as visualization scenarios 1-2. Let us emphasize that the association rules are treated without implication relation. Here, the intervals of

Table 2: Mined numerical features in association rules.

Feature	Scenario 1	Scenario 2
Duration	[76.67,78.07]	[46.95,65.87]
Distance	[14.28,26.32]	[26.24,53.30]
Average HR	[78.79,114.92]	[104.12,141.40]
Average altitude	[631.70,1809.21]	[17.59,547.05]
Calories	[774.92,1161.43]	[1096.82,1209.00]
Ascent	[0.00,10.00]	[0.00,74.19]
Descent	[0.00,54.19]	[0.00,623.88]

numerical attributes are important for the proposed visualization method only. On the other hand, the proposed approach is not limited by a huge number of triangle diagrams, because here we are focused on visualization of the best association rules according to some criteria that normally have a limited number of numerical attributes.

Experimental figures were drawn using the Matlab software framework, using the colored 3-D ribbon plot. All the Figures start at $Location = 0$ and spread on the x axis. The height of the triangles is symbolized at the z axis, and the shades of color are represented by a vertical color-bar. On the other hand, the y axis does not include any meaning.

The visualization of the two mentioned scenarios are presented in the remainder of the paper.

4.1 Scenario 1

The best association rule is presented in Table 3, where the feature F-2 (i.e., "Distance") is compared with the closest features according to a confidence, i.e., "Average HR", "Average altitude", "Descent", "Duration", "Ascent", "Calories".

Table 3: Scenario 1 in numbers.

Scenario	$supp(X)$	$conf(X \Rightarrow Y_i)$					
1	F-2	F-3	F-4	F-8	F-1	F-7	F-6
	0.40	0.19	0.16	0.06	0.05	0.02	0.01

The corresponding visualization of these data are illustrated in Fig. 3, from which it can be seen that the feature "Distance" has higher interdependence with features "Average HR" and "Average altitude" only, but the relationships among the other features do not have a higher effect on the performance of the athlete.

4.2 Scenario 2

In this scenario, only six features are incorporated, because the feature R-6 (i.e., "Calories") does not affect the performance of the cyclist in training (Table 4).

Table 4: Scenario 2 in numbers.

Scenario	$supp(X)$	$conf(X \Rightarrow Y_i)$					
2	F-8	F-3	F-4	F-1	F-7	F-2	F-6
	0.93	0.85	0.75	0.57	0.26	0.22	0.00

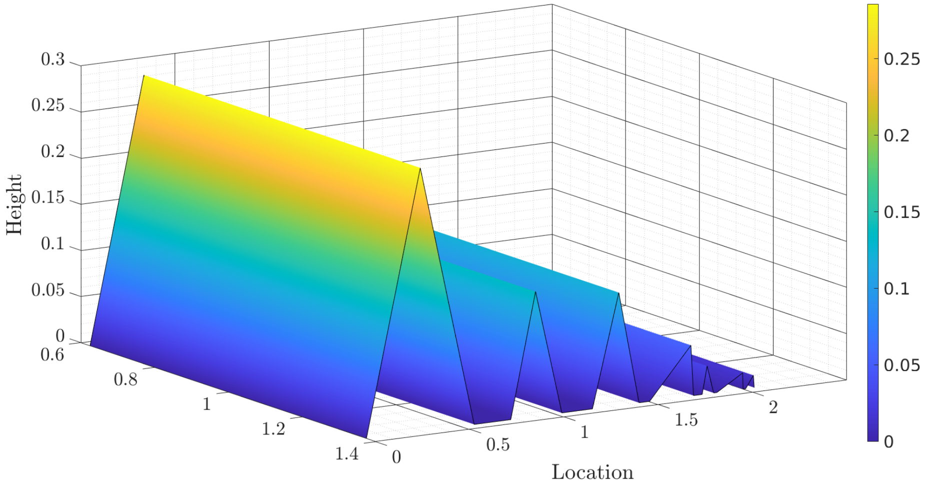

Fig. 3: Visualization of Scenario 1.

As can be seen from Fig. 4, there are six hills, where the former three hills are comparable with the first one according to the area, while the last two expose the lower interdependence. Indeed, the higher interdependence is also confirmed by the larger distances between hills.

5 Conclusion

ARM using numerical attributes of features was rarely applied in practice. The task of the algorithm for ARM with numerical attributes is to find the proper boundary values of numerical features. Consequently, these values specify mined association rules using different values of support and confidence. Thus, the association rules with the best values of support and their closeness to the other features are the more interesting for the user. The users suffer from the lack of information that is hidden in association rules. Obviously, the solution of the problem presents various visualization tolls for extracting the knowledge hidden in data.

This paper proposes a new visualization method inspired by the TDF. Similar as in the TDF cycle race, where the hill climbers have more chances to win the race when the race contains more hills to overcome, the virtual hill slopes, reflecting a probability of one attribute to be more interesting than the other, help a user to understand the relationships among attributes in a selected association rule.

Thus, the relationships between features in the transaction database are illustrated using triangles, representing hills, that need to be overcome by the cyclists. The first triangle in a sequence is a model, because it contains the largest area. The other triangles represent the opponents in the following sense:

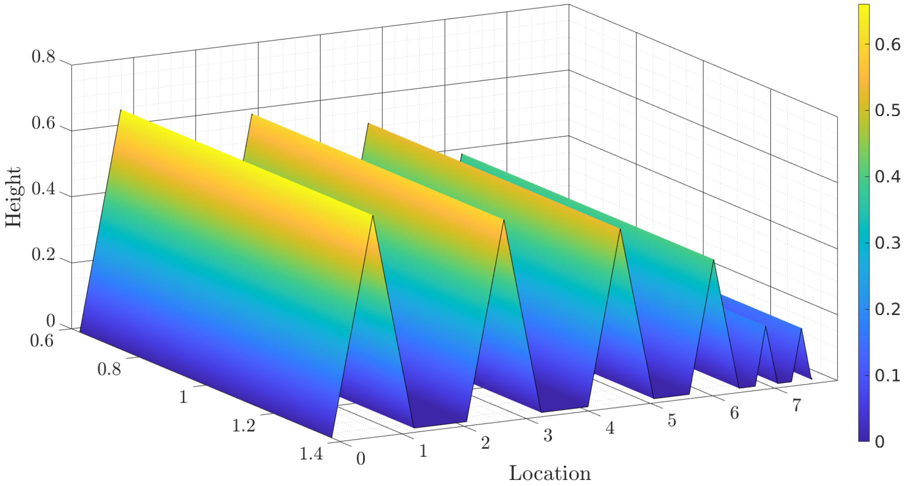

Fig. 4: Second visualization.

The larger the area of a definite triangle, the easier it is for the opponent to overcome. In the ARM sense, this means the following: The larger the triangle, the closer the feature in the transaction database.

The visualization method was employed on a transaction database consisting of features characterizing the realized sports training sessions. Two scenarios were visualized, based on two selected mined association rules. The results of visualization showed the potential of the method, that is able to illustrate the hidden relationships in a transaction database in an easy and understandable way to the user.

In the future, the method could also be broadened for dealing with mixed attributes, i.e., numerical and categorical. The method should be applied to another transaction databases.

Acknowledgements

Iztok Fister thanks the financial support from the Slovenian Research Agency (Research Core Funding No. P2-0042 - Digital twin). Iztok Fister Jr. thanks the financial support from the Slovenian Research Agency (Research Core Funding No. P2-0057). Dušan Fister thanks the financial support from the Slovenian Research Agency (Research Core Funding No. P5-0027). J. Del Ser and E. Osaba would like to thank the Basque Government through EMAITEK and ELKARTEK (ref. 3KIA) funding grants. J. Del Ser also acknowledges funding support from the Department of Education of the Basque Government (Consolidated Research Group MATHMODE, IT1294-19). Andres Iglesias and Akemi Galvez acknowledge financial support from the project PDE-GIR of the European Unions Horizon 2020 research and innovation programme under the Marie

Sklodowska-Curie grant agreement No 778035, and the Spanish Ministry of Science, Innovation, and Universities (Computer Science National Program) under grant #TIN2017-89275-R of the Agencia Estatal de Investigacin and European Funds EFRD (AEI/FEDER, UE).

References

1. R. Agrawal, T. Imieliński, and A. Swami. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, SIGMOD '93, pages 207–216, New York, NY, USA, 1993. ACM.
2. R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *Proc. of 20th Intl. Conf. on VLDB*, pages 487–499, 1994.
3. E. V. Altay and B. Alatas. Performance analysis of multi-objective artificial intelligence optimization algorithms in numerical association rule mining. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–21, 2019.
4. A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115, 2020.
5. P. S. Fader, B. G. S. Hardie, and J. Shang. Customer-base analysis in a discrete-time noncontractual setting. *Marketing Science*, 29, 2010.
6. I. Fister Jr., A. Iglesias, A. Galvez, J. Del Ser, E. Osaba, and I. Fister. Differential evolution for association rule mining using categorical and numerical attributes. In *International Conference on Intelligent Data Engineering and Automated Learning*, pages 79–88, 2018.
7. M. Hahsler and R. Karpienko. Visualizing association rules in hierachical groups. *Journal of Business Economics*, 87:317–335, 2017.
8. J. Han, J. Pei, and Y. Yin. Mining Frequent Patterns without Candidate Generation. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, SIGMOD '00, pages 1–12, New York, NY, USA, 2000. Association for Computing Machinery.
9. T. Y. Lee and E. T. Bradlow. Automated marketing research using online customer reviews. *Journal of Marketing Research*, 48(5):881–894, 2011.
10. A. Lucía, C. Earnest, and C. Arribas. The Tour de France: A physiological review, 2003.
11. A. Lucía, J. Hoyos, A. Santalla, C. Earnest, and J. L. Chicharro. Tour de France versus Vuelta a España: Which is harder? *Medicine and Science in Sports and Exercise*, 35(5):872–878, 2003.
12. O. Netzer, R. Feldman, J. Goldenberg, and M. Fresko. Mine your own business: Market-structure surveillance through text mining. *Marketing Science*, 31, 2012.
13. N. Rogge, D. V. Reeth, and T. V. Puyenbroeck. Performance evaluation of tour de france cycling teams using data envelopment analysis. *International Journal of Sport Finance*, 8(3):236–257, 2013.
14. R. P. Rooderkerk, H. J. Van Heerde, and T. H. Bijmolt. Optimizing retail assortments. *Marketing Science*, 32(5):699–715, 2013.
15. D. Sanders and M. Heijboer. Physical demands and power profile of different stage types within a cycling grand tour. *European Journal of Sport Science*, 19(6):736–744, 2019.

16. A. Santalla, C. P. Earnest, J. A. Marroyo, and A. Lucía. The Tour de France: An updated physiological review, 2012.
17. T. A. Sundhagen. Lance Armstrong: an American legend?, 2011.
18. B. Torgler. La Grande Boucle: Determinants of Success at the Tour de France. *Journal of Sports Economics*, 8(3):317–331, 2007.
19. T. Van Erp, M. Hoozemans, C. Foster, and J. J. De Koning. Case Report: Load, Intensity, and Performance Characteristics in Multiple Grand Tours. *Medicine and Science in Sports and Exercise*, 52(4):868–875, 2020.
20. M. J. Zaki, S. Parthasarathy, M. Ogihara, and W. Li. New Algorithms for Fast Discovery of Association Rules. Technical report, USA, 1997.