Ann. Data. Sci. (2015) 2(2):145–155
DOI 10.1007/s40745-015-0038-8

Identifying High-Number-Cluster Structures in RFID Ski Lift Gates Entrance Data

Boris Delibašić[1] · **Zoran Obradović**[2]

Received: 25 June 2015 / Revised: 27 June 2015 / Accepted: 29 June 2015 / Published online: 9 July 2015

Abstract In this paper we identify skier groups in data from RFID ski lift gates entrances. The ski lift gates' entrances are real-life data covering a 5-year period from the largest Serbian skiing resort with a 32,000 skier per hour ski lift capacity. We utilize three representative algorithms from three most widely used clustering algorithm families (representative-based, hierarchical, and density based) and produce 40 algorithm settings for clustering skiing groups. Ski pass sales data was used to validate the produced clustering models. It was assumed that persons who bought ski tickets together are more likely to ski together. AMI and ARI clustering validation measures are reported for each model. In addition, the applicability of the proposed models was evaluated for ski injury prevention. Each clustering model was tested on whether skiing in groups increases risk of injury. Hierarchical clustering algorithms showed to be very efficient in terms of finding the high-number-cluster structure (skiing groups) and for detecting models suitable for injury prevention. Most of the tested clustering algorithms models supported the hypothesis that skiing in groups increases risk of injury.

Keywords Skiing groups · Ski lift gates RFID data · Hierarchical clustering · K-means · OPTICS · Ski injury

✉ Boris Delibašić
boris.delibasic@fon.bg.ac.rs

Zoran Obradović
zoran.obradovic@temple.edu

1 Faculty of Organizational Sciences, University of Belgrade,
154 Jove Ilića St., Belgrade, Serbia

2 Center for Data Analytics and Biomedical Informatics, Temple University,
1925 N. 12th Street (SERC 035-02), Philadelphia, PA 19122-1801, USA

1 Introduction

Clustering algorithms that are automatically looking for the right number of clusters in data tend to detect fewer clusters, so a high number cluster structure is hard to reveal [20]. These algorithms work soundly in finding k clusters in data when k (number of clusters) is much smaller than n (number of objects). When this is not satisfied, clustering algorithms have difficulties in identifying the hidden high-number-clustering structure. These high-number-clustering structures can be often found in real-life applications (e.g. disease prediction with microarray data [20], clustering of human activity patterns [9]).

Extracting knowledge from sensor data (sensor data mining) is a new and important research field in the big data research field (e.g. [1]). Sensor data is available more and more and due to the large amount of this data there is a huge area of analysis for this kind of data. Sensor data comes from wireless sensor networks, sensor streams, sensor networks, mobile objects, RFID tags and similar.

RFID data is playing a more and more important role in our lives. How to analyze and discover knowledge from RFID data sets is an urgent and challenging research field [7]. Although most of the literature employs hierarchical clustering to find natural groups in data [9], D'Urso and Massari [9] propose a fuzzy approach for clustering path data, since sequences of human activities are typically characterized by switching behaviors, which are likely to produce overlapping clusters. They use two modifications of the fuzzy c-medoids algorithms to cluster human path data. Lv et al. [15] claim that trajectory clustering is usually performed with three approaches: partioning (k-means) clustering, density-based clustering and time-based clustering. The same authors modeled mobile users similarity based on a proposed hierarchical clustering algorithm that uses the cosine distance for measuring similarity.

Based on the literature review on clustering RFID data, in this paper the most frequently used clustering approaches and their representatives are used, i.e. K-means [12], hierarchical clustering [12], and OPTICS [2] as representatives of three clustering algorithm families (representative-based, hierarchical, and density based). The three algorithms were set up (with varying similarity measures, and stopping criteria) so they produce 40 different algorithm settings for clustering high-number-cluster structures.

The 40 algorithm settings were applied on a large real-life ski lift gates entrance dataset covering a five year period. A potential application of the produced models was shown for the case of ski injury prevention. The question of whether skiing in groups (i.e. $|\text{cluster}| > 1$) increases risk of injury was tested.

This paper makes a twofold contribution. On the one hand it proposes algorithm settings that can be used for mining high-number-cluster structures (here skiing groups) for real-life applications. On the other hand it contributes to ski injury prevention, as currently there is no research on the influence of skiing groups on ski injury, and methods for identifying groups from RFID ski lift gate date are missing.

The rest of the paper is structured as follows: In Sect. 2 the big data for analysis are presented. Section 3 explains the algorithms and their settings to analyze the data. Section 4 presents a background on ski injury research, for which the results of the clustering models could be valuable. Section 5 discusses the results. The conclusion of the paper and directions for further research are given in Sect. 6.

2 The Data

The data is from the largest Serbian ski resort, Mt. Kopaonik. The data spans five consecutive seasons (2006–2010). Regulations on Mt. Kopaonik are that each person must buy a ski pass in order to use ski lifts. The Radio-frequency identification (RFID, i.e. wireless non-contact use of radio-frequency electromagnetic fields to transfer data, for the purposes of automatically identifying and tracking tags attached to objects) ski pass is used each time a person wants to enter a ski lift through a ski lift gate. Therefore, for all skiers (in this papers skiers will be used as a generic term for all persons using ski lift gates to enter ski lifts, i.e. skiers, snowboarders, etc.), motion data is collected on ski lift gates and stored in the central database.

Databases used in this research are the ski lift gates' entrance database with spatio-temporal data from skiers' movements, ski patrol injury records, and ski pass sales data. The following attributes were used for the analysis:

1. From the ski lift gates database:
   - Ski pass id,
   - Ski lift entered using a ski gate, and
   - Date and time of entering the ski gate.
2. From the ski patrol injury records:
   - Ski pass id from injured skier.
3. From the ski pass sales database:
   - Sales transaction id,
   - Date and time of sales transaction,
   - Ski pass id(i) (i=1,…,m where m is the number of ski passes sold in one transaction).

A total of 109,553 skiers with weekly ski passes (6 or 7 days valid), for a total of 504,749 skier-days and over six million ski lift transportations, were analyzed. There were in total 735 all-type injuries reported. Weekly ski passes were analyzed, as injury records for this subgroup were available. Injury occurrence measures are: injury per thousand skier-days IPTSD = 1.46, and mean days between injury MDBI = 686.73, which is similar to other ski resorts [17]. Ski injuries are rare events usually occurring at small rates (less than 0.2 %).

Skiers' movement data based on ski lift gates is rarely used in ski injury analysis, although there are some papers that use this data [9]. In [13], authors try to predict the number of skier-days. RFID data have been analyzed for bicycle renting habits in Barcelona [11] and London metro commuting patterns [14]. These papers have performed clustering on the sensor data; however they did not try to reveal high-number-clustering structures, i.e. they did not try to detect bicycle drivers or metro users driving or commuting together. D'Urso and Massari [9] analyzed only one day of ski lift gate data but looked for two clusters in data representing loyal (tend to stick to one ski lift) and variety seeking skiers (tend to change ski lifts often).

3 Methods

Three clustering algorithms and 40 algorithm settings were used for clustering the ski lift gates entrance dataset. These are k-means, hierarchical clustering, and OPTICS.

3.1 K-means Clustering

K-means is probably one of the most popular clustering algorithms. It was used often for clustering high-number cluster data [20], and for RFID clustering [15].

The algorithm consists of the following steps:

1. Randomly initialize k representatives. i.e. centroids.
2. Calculate distance from all objects to centroids, and assign each object to it's closest centroid.
3. Recalculate centroids as mean value of all objects assigned to that centroid.
4. Repeat steps 2 and 3 until convergence or another stopping criterion is satisfied.

In order to overcome a wrongly chosen set of centroids in Step 1, we used *online* updating of centroids, i.e. each object is assigned one after the other to a centroid and each assignment influences a centroid's recalculation. This process of assigning objects to centroids is repeated several times until convergence is reached.

We produced 18 K-means cluster algorithm settings by combining

- nine stopping criteria $k = n \times \{10, 20, 30, 40, 50, 60, 70, 80, 90\ \%\}$
- and two distance measures, i.e. cosine distance measure and Euclidean distance measure.

K-means is abbreviated as KM (Table 1), distance measures with "Cos" and "Euc" while numbers 10 through 90 signify the percentage by which the total number of skiers was multiplied to produce k.

3.2 Hierarchical Agglomerative Clustering

Hierarchical agglomerative clustering is a procedure that at each merges an object into a cluster, until all objects are in one cluster. The following steps are used in hierarchical clustering:

1. Calculate an $n \times n$ distance matrix.
2. Merge an object to a cluster or another object according a minimum linkage criterion.
3. Repeat 2 until all objects have merged.

For the linkage criterion we used mean linkage clustering (Eq. 1)

$$\frac{1}{|A|\ |B|} \sum_{a \in A} \sum_{b \in B} d\,(a, b) \tag{1}$$

Table 1 AMI and ARI (mean, rank, and standard deviation), odds ratio, and 95 % confidence interval for 40 algorithm settings

Algorithm	AMI [m (r), s]		ARI [m (r), s]		OR	95 % CI	
HAnd10	0.2264 (19)	0.0328	0.0796 (31)	0.0466	2.393	1.8498	3.0956
HAnd20	0.2862 (1)	0.0396	0.1196 (18)	0.052	1.4422	1.2232	1.7005
HAnd30	0.2841 (2)	0.0462	0.1397 (4)	0.0515	1.2393	1.1013	1.3945
HAnd40	0.27 (5)	0.0501	0.1477 (1)	0.0585	1.2121	1.1057	1.3289
HAnd50	0.2464 (12)	0.0472	0.144 (2)	0.0634	1.244	1.151	1.3445
HAnd60	0.2115 (23)	0.0425	0.128 (11)	0.0613	1.2041	1.1209	1.2934
HAnd70	0.1661 (29)	0.0321	0.1011 (24)	0.0525	1.1972	1.1159	1.2843
HAnd80	0.111 (33)	0.0222	0.0672 (33)	0.036	1.2185	1.1296	1.3144
HAnd90	0.0535 (40)	0.0113	0.0319 (40)	0.0179	1.1927	1.0832	1.3132
HAndSC	0.2487 (10)	0.0485	0.1439 (3)	0.0619	1.1780	1.0868	1.2769
HCos10	0.2196 (22)	0.0307	0.0796 (32)	0.0454	1.249	0.8756	1.7816
HCos20	0.2726 (3)	0.0358	0.1118 (21)	0.0409	1.2576	1.0626	1.4883
HCos30	0.2711 (4)	0.0439	0.1309 (9)	0.0492	1.1084	0.9839	1.2488
HCos40	0.2568 (7)	0.0478	0.1379 (5)	0.0544	1.1381	1.0379	1.2479
HCos50	0.2345 (16)	0.0442	0.1349 (7)	0.0582	1.1877	1.0986	1.2839
HCos60	0.2034 (26)	0.041	0.1218 (16)	0.0589	1.1618	1.0814	1.2483
HCos70	0.162 (30)	0.0324	0.0985 (26)	0.0513	1.1704	1.091	1.2556
HCos80	0.1108 (34)	0.0218	0.0672 (34)	0.0359	1.1912	1.1046	1.2846
HCos90	0.0549 (39)	0.0108	0.0333 (38)	0.018	1.1985	1.0877	1.3205
HCosSC	0.2359 (13)	0.0481	0.1359 (6)	0.06	1.1574	1.0687	1.2534
KMCos10	0.2352 (14)	0.0277	0.0984 (27)	0.026	NA	NA	NA
KMCos20	0.2593 (6)	0.0351	0.124 (14)	0.0395	1.0171	0.4554	2.2719
KMCos30	0.252 (9)	0.0414	0.1312 (8)	0.0495	1.1644	0.9391	1.4439
KMCos40	0.2346 (15)	0.0423	0.1278 (12)	0.0544	1.188	1.0602	1.3313
KMCos50	0.2096 (24)	0.0401	0.1177 (19)	0.0542	1.1003	1.0096	1.1992
KMCos60	0.1809 (27)	0.037	0.1038 (23)	0.0538	1.1316	1.0514	1.2179
KMCos70	0.1475 (31)	0.0283	0.0858 (28)	0.0432	1.1232	1.0472	1.2048
KMCos80	0.1087 (35)	0.0208	0.0648 (35)	0.0346	1.2478	1.157	1.3457
KMCos90	0.0623 (37)	0.012	0.0366 (37)	0.0197	1.1984	1.0946	1.3119
KMEuc10	0.2198 (21)	0.031	0.0827 (30)	0.0242	0.7341	0.3491	1.5439
KMEuc20	0.254 (8)	0.0325	0.1112 (22)	0.0328	0.8996	0.6816	1.1874
KMEuc30	0.2472 (11)	0.0382	0.1219 (15)	0.0414	0.8277	0.7003	0.9782
KMEuc40	0.2293 (18)	0.0399	0.1214 (17)	0.0491	0.7759	0.6889	0.8739
KMEuc50	0.2053 (25)	0.037	0.1134 (20)	0.0515	0.8604	0.7869	0.9407
KMEuc60	0.1768 (28)	0.034	0.1007 (25)	0.0497	0.9213	0.8544	0.9935
KMEuc70	0.145 (32)	0.028	0.0844 (29)	0.0429	0.9768	0.9105	1.0478
KMEuc80	0.1075 (36)	0.02	0.064 (36)	0.0335	1.1775	1.0925	1.2692
KMEuc90	0.0551 (38)	0.0111	0.0325 (39)	0.0184	1.249	1.1334	1.3764
OPTAnd	0.2327 (17)	0.0437	0.1303 (10)	0.059	1.1523	1.0583	1.2546
OPTCos	0.2249 (20)	0.0408	0.1251 (13)	0.0561	1.2277	1.1296	1.3344

where A and B are objects or clusters, and d(a, b) are distances between each object a in A and b in B.

Hierarchical agglomerative clustering was used to produced 20 algorithm settings by combining

- nine stopping criteria $k = n \times \{10, 20, 30, 40, 50, 60, 70, 80, 90\ \%\}$, and
- two distance measures, i.e. cosine distance measure and absolute normalized difference "And" [14] which produces 18 algorithm settings.

In addition, for the remaining two algorithm settings, we also propose a stopping criterion for hierarchical clustering that can identify high-number-cluster structures. The following criterion is proposed in this paper: *Maximize the number of non-single skiing clusters max* $\left(\sum |cluster > 1|\right)$. This criterion makes a balance between cutting the hierarchical tree too high or too low in the hierarchy. By cutting the tree too high there would be too few clusters, clusters of larger size, and very few single-person clusters. By cutting the tree too low there would be too many single-skier clusters. The proposed stopping criterion makes a compromise because it is intended to produce fewer single-person clusters while keeping the clusters at small sizes.

In total 20 hierarchical algorithm settings were produced. They are shown in Table 1 and begin with an H. Similarity measure abbreviations are "Cos" and "And", while numbers 10 through 90 signify the percentage of n that is used to determine k. The proposed stopping criterion is marked as SC.

3.3 OPTICS

Ordering points to identify a clustering structure (OPTICS) [2] is a density based algorithm that improves the famous DBSCAN [10] clustering algorithm as it can handle neighborhoods with varying densities, which was one of the drawbacks of DBSCAN.

The ε distance is the largest distance considered for clusters. Clusters can be extracted for all ε_i values smaller or equal than ε, which was not possible in DBSCAN. DBSCAN had the ability to find clusters for a specific value of ε, however all sub-clusters (hierarchically nested) within an ε neighborhood cluster would not be identified.

Key notions in OPTICS are the following:

- An object p is in the ε-neighborhood of q if the distance from p to q is less than ε.
- A core object has at least *minPts* in its ε-neighborhood.
- An object p is directly density reachable from object q if q is a core object and p is in the ε-neighborhood of q.
- The reachability-distance of p is the smallest distance such that p is density reachable from a core object o, however it can't be smaller than the core distance of o.
- The core-distance is the smallest distance ε' between p and an object in its ε-neighborhood such that p would be a core object.
- A steep downward point is a point that is lower than its successor by a certain percentage. A steep upward point is similarly defined.

- A steep downward area is a region [a, b] such that a and b are both steep downward points, each successive point is at least as low as its predecessors, and the region does not contain more than *minPts* successive points that are not steep downward.

The definition of a cluster in OPTICS is:

- Starts with a steep downward area,
- Ends with a steep upward area,
- Contains at least minPts,
- The reachability values in the cluster are at least by a certain percentage lower than the first point in the cluster.

In general, OPTICS does not produce a clustering model, but rather an ordering of points with reachability values who determine the closeness of a point to its predecessors. Ankerst et al. (1999) propose a procedure that can be used for automatic cluster extractions from the ordering of points.

```
1 ExtractClusters (orderedObjects);
2 for each object;
3 if start of downward area D;
4 add object to downward areas;
5 elseif start of upward area U;
6 add object to upward areas;
7 for each downward area D;
8 if D and U form a cluster;
9 add [start(D), end(U)] to set of clusters.
```

OPTICS was used to produce two algorithm settings:

– with “And” and “Cos” distance measures.

OPTICS is a density based algorithm that can identify varying density areas. The original algorithm had to be modified to work with this dataset, as the original algorithm was unable to find clusters with size 2 (The definition of a cluster OPTICS allows for a minimum of three objects in a cluster, otherwise the clusters are treated as outliers). We relaxed the conditions of OPTICS to allow identification of clusters sized 2. We fixed the epsilon parameter of OPTICS to a “small” value [2] so all clusters, even the less significant, could be detected. “High” epsilon values would detect only the most significant clusters. A fine tuning of this parameter was left for further research.

Clusters were produced for each day. The following procedure was used to determine clusters:

1. Each skiers’ movement data was recorded in a 168-attribute row. Each row represented a count of how many times a skier during a single day (skier-day) entered a ski lift in a given time frame. As there were in total 21 ski lifts and 8 time bands (one-hour Sakoe-Chiba band [18] from 9 AM to 5 PM), a total of 168 attributes were used for representing each skier’s movement.
2. Clustering models were produced for each skiing day (in total 376 days) using 40 algorithm settings. In total, there were 15,040 clustering models produced.

3. All cluster models were tested with AMI and ARI cluster validation measures, and the results were aggregated for each algorithm setting. The results are shown in Table 1.

Cluster quality was measured with external validation measures i.e. Adjusted Mutual Information (AMI) [19] and Adjusted Rand Index (ARI) [16]. Both indexes compare two clustering models [one from algorithm settings (in total there are 15,040 models) and the other ground truth model based on the sales transactions database, where each transaction (skiers that bought ski passes together) is a cluster]. AMI and ARI value of 1 represent perfect clustering, where 0 would mean that there is no matching between the ground truth cluster model and the clustering model at testing. All values above 0 show evidence about the strength of correlation between the clustering and ground truth.

ARI is a well-known measure of cluster validity, where AMI was recently proposed as a "general purpose" measure for clustering validation which identifies the true number of clusters better than other measures, including ARI.

Please note that sales transaction data cannot be looked at as a "golden" standard for evidence of skiing in groups, but rather as a "silver" standard, because sales transactions data can be only an indicator of skiing together, but not as firm evidence.

4 Background in Ski Injury Research

There is evidence from literature that injured skiers usually (in 87 % of cases) ski in groups [6]. Still, methods for identifying skiing groups, based on similar trajectories, are rarely found in ski injury literature. It is also not clear whether and how skiing groups influence ski injury. To our knowledge, there are currently no papers that test the relationship between groups (size, structure, etc.) and risk of injury. This paper makes a simple test on the relationship between ski groups and injury occurrence, i.e. does skiing in groups influence ski injury. Identification of risk factors and mechanisms that cause injury are necessary in order to reduce injury rates [4].

There are already numerous studies of skier's individual injury risk factor identification. Various risk factors have been reported, such as: gender [17], age [17], personality types [3], skier collision [8], skiing errors [5], speed of skiing [5,8], fatigue [5], perception of low difficulty [5], skillfulness and experience [8], quality of equipment [8], quality of ski slopes and quality of preparation, collision against objects, and jumps [8].

However, all of these risk factors are looking skiers as if they were skiing single. The research on skiing groups and their relation to injury is, to the best of our knowledge, still missing in literature. This paper proposes first steps towards filling in this gap.

5 Results

The results are shown in Table 1. For each algorithm ARI, AMI measures are reported, as well as the odd ratios for each model testing the hypothesis that skiing in groups influences ski injury. The best algorithm according to AMI is HAnd20, while ARI

recommends HAnd40. Both of these best algorithms are hierarchical and use the "And" similarity measure. It can also be noticed that other algorithms recommending the number of clusters k to be 20–30 % of n have pretty sound results according to AMI. ARI recommends k to be 40–50 % and also ranks algorithms that use the proposed stopping criteria in this paper (SC) highly.

The K-means algorithm shows worse results than hierarchical clustering, and using a Euclidean metric (which is the standard metric in the K-means algorithm) is shown to be quite inefficient. For the KMCos10 algorithm setting it was not possible to calculate the odds ratio because the clustering model didn't discriminate between injured and non-injured skiers.

The density based algorithm OPTICS showed medium quality results, but with better adjustment of the epsilon parameter in these two algorithms it is expected that these algorithms could perform better.

When looking at odds ratios, all algorithm models, except KMEuc10 to KMEuc70, recommend that skiing in groups increases the risk of injury. The largest odds ratios are noticed in HAnd10 and HAnd20. HAnd20 reports that odds (risk) of injury are 44, 22 % times greater if skiing in groups. The odds ratio with HAnd10 (2.393) is the highest, which indicates that the clustering model detected with this algorithm would be also applicable for injury prevention.

6 Conclusion

This paper makes a twofold contribution. It proposes methods for identifying high-number-cluster structures and tests models' applicability on ski injury prevention. The clustering results suggest that clustering hidden structures in ski lift data could be useful for finding hidden patterns that can be used for ski injury prevention. It is worth mentioning that most ski resorts in Europe have RFID ski lift gates installed, and could exploit this data for injury prevention. This paper offers several further research possibilities. On the one hand, researchers should analyze cluster models in skier movement data, which would allow for identification of models with high odds ratios (risk factors). This could then be used in detecting and warning skiers that are at greater risk of injury. On the other hand, more clustering algorithms could be used and better adjustment of parameters could be done so the high-number-cluster structures in data could be better revealed. Still, the results of this paper are quite sound, having in mind that the results were tested on "silver truth" standard data. Testing the algorithms on "golden truth" data and on more ski resorts is needed so to provide higher-confidence conclusions. This paper makes first steps towards filling gaps in the literature in revealing high-number-cluster structures in RFID human activity data and in using this revealed structure for ski injury prevention.

Acknowledgments This research is partially funded by the US Department of State CIES Fulbright Visiting Program grant, conducted at the Center for Data Analysis and Biomedical Informatics (DABI) at Temple University. The authors acknowledge the Ski Resorts of Serbia for providing data for this research, and for providing support throughout the research. The authors are also very grateful to the Mountaineer Rescue Service of Serbia (Ski Patrol) for providing data for this analysis.

References

1. Aggarwal CC (2013) Managing and mining sensor data. Springer, Heidelberg
2. Ankerst M, Breunig MM, Kriegel HP, Sander J (1999) Optics: ordering points to identify the clustering structure. ACM Sigmod Record. doi:10.1145/304182.304187
3. Castanier C, Scanff CL, Woodman T (2010) Who takes risks in high-risk sports?. Res Quart Exerc Sport, A typological personality approach. doi:10.1080/02701367.2010.10599709
4. Chalmers DJ (2002) Injury prevention in sport: not yet part of the game? Inj Prev. doi:10.1136/ip.8.suppl_4.iv22
5. Chamarro A, Fernández-Castro J (2009) The perception of causes of accidents in mountain sports: a study based on the experiences of victims. Acc Anal Prev. doi:10.1016/j.aap.2008.10.012
6. Cooper N (2008) Correlative study into injury epidemiology, use of protective equipment and risk taking among adolescent participants in alpine snow sports. J ASTM Int. doi:10.1520/JAI101371
7. Deng H, Lin G (2009) PDSC: Clustering object paths from RFID data sets. In information processing, in IEEE asia-pacific conference, 2:541–544
8. Dohin B, Kohler R (2008) Traumatologie du ski et du snowboard chez l'enfant et l'adolescent: épidémiologie, physiopathologie, prévention et principales lésions. Arch Pediatr. doi:10.1016/j.arcped.2008.08.022
9. D'Urso P, Massari R (2013) Fuzzy clustering of human activity patterns. Fuzzy Set Syst. doi:10.1016/j.fss.2012.05.009
10. Ester M, Kriegel HP, Sander J, Xu X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the 2nd international conference on knowledge discovery and data mining, AAAI Press, Portland, OR, 226–231
11. Froehlich J, Neumann J, Oliver N (2009) Sensing and predicting the pulse of the city through shared bicycling. IJCAI 9:1420–1426
12. Hastie T, Tibshirani R, Friedman JJH (2001) The elements of statistical learning, vol 1. Springer, New York
13. King MA, Abrahams AS, Ragsdale CT (2014) Ensemble methods for advanced skier days prediction. Expert Syst Appl 41(4):1176–1188
14. Lathia N, Smith C, Froehlich J, Capra L (2013) Individuals among commuters: building personalised transport information services from fare collection systems. Pervasive Mobile Comput. doi:10.1016/j.pmcj.2012.10.007
15. Lv M, Chen L, Chen G (2013) Mining user similarity based on routine activities. Inf Sci 236:17–32
16. Milligan GW, Cooper MC (1987) Methodology review: clustering methods. App Psych Meas 11(4):329–354
17. Ruedl G, Kopp M, Sommersacher R, Woldrich T, Burtscher M (2013) Factors associated with injuries occurred on slope intersections and in snow parks compared to on-slope injuries. Accid Anal Prev. doi:10.1016/j.aap.2012.09.019
18. Sakoe H, Chiba S (1978) Dynamic programming algorithm optimization for spoken word recognition. IEEE Trans Acoust Speech Sig Proc, V ASSP 26:43–49
19. Vinh NX, Epps J, Bailey J (2010) Information theoretic measures for clusterings comparison: variants, properties, normalization and correction for chance. J Mach Learn Res 11:2837–2854
20. Vukićević M, Kirchner K, Delibašić B, Jovanović M, Ruhland J, Suknović M (2012) Finding best algorithmic components for clustering microarray data. Knowl Inf Syst. doi:10.1007/s10115-012-0542-5

Boris Delibašić is associate professor at the University of Belgrade, Faculty of Organizational Sciences (School of Management). In 2007 he received his Ph.D. from the University of Belgrade. From 2012 he holds his current position. Since 2011 he is coordination board assistant at the EURO working group on Decision Support Systems (EWG-DSS). His main research interests are decision support systems, machine learning algorithm design, business intelligence, and multi-attribute decision making. Dr. Delibašić research profile can be found here.

Zoran Obradović is a L.H. Carnell Professor of Data Analytics at Temple University, Professor in the Department of Computer and Information Sciences with a secondary appointment in Statistics, and is the Director of the Center for Data Analytics and Biomedical Informatics. His research interests include data mining and complex networks applications in health management and other complex decision support systems. He is the executive editor at the journal on Statistical Analysis and Data Mining, which is the official publication of the American Statistical Association and is an editorial board member at eleven journals. He is the chair at the SIAM Activity Group on Data Mining and Analytics and was co-chair for 2013 and 2014 SIAM International Conference on Data Mining and was the program or track chair at many data mining and biomedical informatics conferences. His work is published in more than 300 articles and is cited more than 15,000 times (H-index 48). For more details see http://www.dabi.temple.edu/~zoran/.