Application of the DBSCAN method in analyzing aerobic and anaerobic thresholds in athletes – A new approach to data segmentation in sports training

Dariusz Kaczmarek, Michał Hamela and Marcin Mokrzycki

Politechnika Opolska, Prószkowska 76, Opole, 45-758, Poland

ARTICLE INFO

Keywords:
DBSCAN
Aerobic Threshold
Anaerobic Threshold
Data Clustering
Athlete Performance
Density-Based Clustering

ABSTRACT

This article explores the application of the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) method in analyzing aerobic and anaerobic thresholds in athletes. Identifying these physiological thresholds is essential for optimizing training programs and improving athletic performance. Traditional approaches often rely on predetermined threshold values, which may not accurately reflect the diverse physiological profiles of athletes. DBSCAN, a clustering algorithm that groups data points based on density, offers a data-driven approach to segmenting threshold data without prior assumptions. By leveraging this method, we can better identify unique patterns in oxygen and lactate levels, providing more personalized and adaptive training insights. This approach holds promise for enhancing training efficacy and tailoring programs to individual athletes' needs, thereby advancing the field of sports science.

1. Introduction

Lactate analysis has long been central to sports science, providing insights into an athlete's metabolic responses under varying exercise intensities (1; 2). For over two centuries, researchers have explored lactate's role in both anaerobic and aerobic glucose metabolism, generating ongoing discussions around its measurement and interpretation. Concepts such as the lactate threshold, lactate turn point, onset of blood lactate accumulation (OBLA), maximal lactate steady state (MLSS), anaerobic threshold, ventilatory threshold, and individual anaerobic threshold have introduced valuable, yet sometimes confusing, metrics for performance assessment (3). These thresholds underpin a tri-phasic model of energy delivery and lactate production, distinguishing aerobic (AeT) and anaerobic thresholds (AnT), which are critical for assessing an athlete's endurance and overall performance (4).

The aerobic and anaerobic thresholds are key markers in evaluating an athlete's capacity and performance level. They offer insights into physiological capabilities, highlighting an athlete's ability to manage and clear lactate, a byproduct of intense exercise that can contribute to fatigue. Higher AeT and AnT values indicate a more efficient physiological system, often correlating with greater endurance. Athletes with elevated thresholds can sustain high-intensity exercise longer without significant fatigue, a crucial factor in endurance sports like long-distance running, cycling, or triathlons (1; 2; 5; 6; 7).

AeT and AnT are connected to different energy systems, and understanding these thresholds helps athletes target the appropriate energy systems for various types of sports. Training within AeT and AnT zones aligns exercise intensity with the athlete's physiological capacity, leading to more efficient and targeted training outcomes (1). Athletes and coaches use these thresholds to fine-tune training intensities and prevent overtraining, which can lead to burnout and potential injuries if unmanaged.

Moreover, knowledge of AeT and AnT can guide tactical decisions in team sports. Coaches may use this information to make strategic substitutions or adjust tactics based on the athlete's current energy levels (8; 9). Recognizing the signs of overtraining, as indicated by AeT and AnT data, enables immediate intervention, such as reducing training intensity, incorporating rest days, or modifying the training program to allow recovery (10).

In recent years, machine learning (ML) techniques have become instrumental in predictive analysis across multiple domains, including sports science (11). Among these techniques, the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm has shown promise for identifying patterns in complex, multidimensional datasets. In the context of sports performance, DBSCAN can be effectively used to cluster physiological data, such as heart rate, oxygen consumption, and blood lactate levels, allowing for the identification of distinct performance zones corresponding to the aerobic threshold (AeT) and anaerobic threshold (AnT) (12; 13; 14). By grouping similar data points and distinguishing outliers, DBSCAN provides a robust approach for analyzing the variability and transitions between these thresholds.

This study aims to assess the potential of DBSCAN in accurately identifying AeT and AnT by clustering physiological data and comparing the resulting clusters with thresholds determined by conventional methods. Utilizing DBSCAN's capacity to reveal underlying patterns in physiological responses, we seek to enhance our understanding of these thresholds, thereby supporting a more individualized approach to athlete performance and endurance training.

2. Objective of the work

The aim of this study is to conduct a comprehensive analysis of data concerning the aerobic and anaerobic thresholds of athletes using the DBSCAN clustering method. The key aspects of the study include:

2.1. Understanding and identifying patterns

- Investigating how various factors, such as the type of sport, training level, age, and gender, influence athletes' aerobic and anaerobic thresholds.
- Analyzing the data will allow for the identification of specific patterns that may indicate different training strategies or potential areas for improving performance.

2.2. Application of the DBSCAN Algorithm

- Utilizing the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm for clustering the data. DBSCAN is particularly useful in situations where the data is heterogeneous and contains noise, which is typical for sports data.
- Careful configuration of the algorithm's parameters, such as the neighborhood radius (epsilon) and the minimum number of points (minPts) in a cluster, to achieve optimal clustering results.

2.3. Analysis of results

- Conducting a detailed analysis of the clustering results to determine the characteristics of the different groups of athletes, as well as the differences and similarities in their aerobic and anaerobic thresholds.
- Visualizing the clusters using appropriate graphical techniques, such as scatter plots, 3D plots, and other data visualization methods, to facilitate the interpretation of the results.

2.4. Training recommendations

- Developing personalized training programs based on the clustering results. These recommendations will take into account the individual needs of athletes in the context of their performance and training goals.
- Investigating how appropriate adjustments to training programs based on identified groups can enhance training effectiveness and contribute to better sports performance.

2.5. Practical application of results

- The study aims not only to theoretically develop a clustering method but also to provide practical tools for coaches and athletes that can be used in daily training and competition preparations.
- The application of the research findings in sports practice may contribute to better planning of training cycles and increase the efficiency of preparatory processes.

2.6. Conclusions and future research directions

- Summarizing the main findings and conclusions resulting from the analysis of data clustering, as well as indicating possible directions for further research in this area.
- Proposing subsequent research steps that could expand the understanding of phenomena related to sports performance and their applications in practice.

3. Methodology

- **Data Collection:** The analysis will be based on the results of fitness tests conducted on a group of athletes. In the absence of access to data, it may be possible to use public data or conduct original tests.
- **Statistical Analysis:** Applying statistical analysis tools for preliminary data processing, including normalization, outlier removal, and exploratory data analysis.
- **Use of Computational Tools:** Implementing the DBSCAN algorithm in Python, utilizing libraries such as scikit-learn and pandas for data analysis and visualization.

4. Expected outcomes

- Enabling a better understanding of the dynamics of athletes' performance based on their results in aerobic and anaerobic tests.
- Developing practical tools and strategies for coaches that may lead to improved sports performance outcomes.
- Creating a foundation for further research into the application of data analysis methods in sports.

This work aims at not only theoretical advancement but also practical application of modern data analysis methods in the field of sports, which may contribute to innovations in training approaches and athlete preparation.

References

[1] Ghosh AK. Anaerobic threshold: its concept and role in endurance sport. *The Malaysian Journal of Medical Sciences: MJMS*. 2004; 11:24–36. PMID: 22977357.

[2] Faude O, Kindermann W, Meyer T. Lactate threshold concepts: how valid are they? *Sports Medicine*. 2009; 39:469–490. PMID: 19453206.

[3] Krishnan A, Guru CS, Sivaraman A, Alwar T, Sharma D, Angrish P. Newer Perspectives in Lactate Threshold Estimation for Endurance Sports—A Mini-Review. *Central European Journal of Sport Sciences and Medicine*. 2021; 35:99–116. https://doi.org/10.18276/cej.2021.3-09.

[4] Binder RK, Wonisch M, Corra U, Cohen-Solal A, Vanhees L, Saner H, et al. Methodological approach to the first and second lactate threshold in incremental cardiopulmonary exercise testing. *European Journal of Cardiovascular Prevention and Rehabilitation*. 2008; 15:726–734. https://doi.org/10.1097/HJR.0b013e328304fed4. PMID: 19050438.

[5] Neufeld EV, Wadowski J, Boland DM, Dolezal BA, Cooper CB. Heart Rate Acquisition and Threshold-Based Training Increases Oxygen Uptake at Metabolic Threshold in Triathletes: A Pilot Study. *International Journal of Exercise Science*. 2019; 12:144–154. PMID: 30761193.

[6] Meyer T, Gabriel HHW, Kindermann W. Is determination of exercise intensities as percentages of OV_{O2max} or HR_{max} adequate? *Medicine and Science in Sports and Exercise*. 1999; 31:1342–1345. https://doi.org/10.1097/00005768-199909000-00017. PMID: 10487378.

[7] Lounana J, Campion F, Noakes TD, Medelli J. Relationship between Medicine and Science in Sports and Exercise. 2007; 39:350–357. https://doi.org/10.1249/01.mss.0000246996.63976.5f. PMID: 17277600.

[8] Piero DWD, Valverde-Esteve T, Castań JCR, Pablos-Abella C, Díaz-Pintado JVSA. Effects of work-interval duration and sport specificity on blood lactate concentration, heart rate and perceptual responses during high intensity interval training. *PLOS ONE*. 2018; 13(7):e0200690. https://doi.org/10.1371/journal.pone.0200690.

[9] Hwang JH, Moon NR, Heine O, Yang WH. The ability of energy recovery in professional soccer players is increased by individualized low-intensity exercise. *PLOS ONE*. 2022; 17(6):e0270484. https://doi.org/10.1371/journal.pone.0270484. PMID: 35771850.

[10] Yu Y, Li D, Lu Y, Jing M. Relationship between methods of monitoring training load and physiological indicators changes during 4 weeks cross-country skiing altitude training. *PLOS ONE*. 2023; 18(12):e0295960. https://doi.org/10.1371/journal.pone.0295960. PMID: 38100499.

[11] Krishnan A, Sivaraman A, Alwar T, Sharma D, Upadhyay V, Kumar A. Relevance of lactate threshold in endurance sports: a review. *European Journal of Pharmaceutical and Medical Research*. 2020; 7:513–524.

[12] Etxegarai U, Portillo E, Irazusta J, Arriandiaga A, Cabanes I. Estimation of lactate threshold with machine learning techniques in recreational runners. *Applied Soft Computing*. 2018; 63:181–196. https://doi.org/10.1016/j.asoc.2017.11.036.

[13] Huang S, Casaburi R, Liao M, Liu K, Chen YJ, Fu T, et al. Noninvasive prediction of Blood Lactate through a machine learning-based approach. *Scientific Reports*. 2019; https://doi.org/10.1038/s41598-019-38698-1. PMID: 30778104.

[14] Govers R. Predicting Heart Rates Of Sport Activities Using Machine Learning;. Available from: http://essay.utwente.nl/85685/1/Govers_BA_EEMCS.pdf.