UNIVERSITÄT
BERN

Biomimicry-based Algorithms and Their Lack of Generalization

Bachelor Thesis

Dean Klopsch

from

Biel BE, Switzerland

Faculty of Science
University of Bern

16 February 2021

Prof. Dr. Oscar Nierstrasz

Pascal Gadient

Software Composition Group

Institute of Computer Science

University of Bern, Switzerland

Abstract

Biomimicry has received much attention in engineering, and many breakthrough discoveries have been guided by a solution found in nature. However, many biomimicry-based proposals apply to a specific problem, provide limited context, and lack implementation details. That makes it unnecessarily hard for practitioners to find relevant literature for their problems.

To investigate this problem, we performed a literature review on 111 publications related to biomimicry and extracted several characteristics, *e.g.*, meta-data, the solution, and the investigated species. In particular, we were interested in whether the proposed algorithms could be used for other use cases.

Our results indicate a structural issue: publications related to new or adapted algorithms very prominently emphasize a specific use case, instead of the generalized problem category, *e.g.*, clustering. We found that 38% lack generalization in at least one of the introductory elements (*i.e.*, title, abstract, and introduction), and that 53% of them lack generalization entirely. Moreover, 40% of the proposed algorithms lack at least one major characteristic, *e.g.*, code samples or benchmarks against state of the art algorithms.

We motivate the generalization problem with our adapted implementation of an algorithm proposed for load scheduling. Moreover, the artifacts of this study can support practitioners in finding more efficiently existing solutions across research domains.

Contents

1 Introduction

We are surrounded by problems and solutions that have been perfected over thousands of years by nature, but unfortunately they are unnecessarily hard to find: While reviewing existing work to find applicable algorithms for a mathematical problem, we realized that essential information was missing in many publications, making it hard for practitioners to fully understand and adapt the proposed approaches.

For example, a paper titled “Hybrid Genetic-Gravitational Search Algorithm for Load Scheduling in Cloud Computing” [2] proposes a new algorithm specifically tailored for load scheduling in cloud computing, however the algorithm itself is solving a clustering problem. Therefore, the very same algorithm can be used to cluster *any* data.

Researchers have investigated ways to interact with unstructured data and proposed semantic search engines [4, 6, 9] and well-thought methodologies [7, 11], but unfortunately, such tools require additional effort from the user, they require expert knowledge to master them, and the methodologies barely consider this aspect.

To investigate the prevalence of that problem, we performed a literature review on papers which relate to biomimicry, and finally collected 111 publications in six categories (algorithm, concept, framework, comparison, survey, review). In these works, we were interested in the used models, similarities among them, and other characteristics.

We motivate our work with an implementation of the aforementioned load scheduling algorithm in which

we show that the proposed algorithm supports another use case than the one that has been reported. This is due to more general nature of the proposed algorithm that has not been covered: its ability to find solutions for combinatorial problems, *e.g.*, clustering.

In order to assess the problem for practitioners, we answer the following research questions:

RQ$_1$: *What biomimicry literature is available?* We found 68 journal articles and 43 conference or workshop paper submissions. Algorithmic submissions dominate (89 submissions, 80%), followed by novel concepts and reviews (each 6, 5%). Each remaining category, *i.e.*, comparison, framework, and survey has been assigned to less than 4% of all submissions. The *People's Republic of China* contributed the most publications (18%), followed by India (12%), Iran (9%), and the United Kingdom (6%). The United States (4%), Australia (3%) and major countries in Western Europe (<=3%) remain rather at the end of the spectrum.

RQ$_2$: *What are the characteristics of the proposed algorithms in the existing literature, and how comprehensive is their support?* When considering publications related to algorithms, only a minority (15, 17%) introduce a novel algorithm. Most publications (51, 57%) try instead to improve meta-heuristics, *e.g.*, by adjusting parameters or introducing additional features. About 26% are dedicated to a specific use case; often adapting an existing concept to an unrelated field of application. 38% lack a discussion of their algorithm's generalization capabilities in at least one of the introductory elements (*i.e.*, title, abstract, introduction), and 53% of them lack generalization entirely. That is, they do not provide any information to readers about other problem domains for which their algorithm could be used. Furthermore, 40% lack at least one major characteristic, *i.e.*, they did not provide any information about the stability, reliability, efficiency, or the pseudo-code of the algorithm.

RQ$_3$: *How can we generalize a specialized algorithm, and what are the gains?*

We implemented a highly specialized algorithm in a different context and show that generalization can be achieved while still maintaining the same benefits. Our implementation is able to cluster candidates more efficiently than its variant without the gravity component, and than other rather trivial approaches.

In summary, we collected 111 publications related to biomimicry and analysed their metadata as well as their proposed algorithms. We adapted an algorithm tailored for a very specific use case and made it suitable for any clustering problem. The supplemental material is available online,[1] *i.e.*, the list of papers including their characteristics, and the source-code of the algorithm that has been adapted for other use cases.

The remainder of this thesis is structured as follows: We present the metadata from the literature review in chapter 2, then we report properties of the found algorithms in chapter 3, and we exemplify the adaptation of a specialized algorithm in chapter 4. Finally, we declare our threats to validity in chapter 5, we list the related work in chapter 6, and we conclude in chapter 7.

[1] `https://github.com/Dean442/BSc-Thesis-Supplement`

2

Literature Review

In this section, we first describe the process we have followed to collect relevant publications, before we answer $\mathbf{RQ}_1$: *What biomimicry literature is available?* based on the metadata of the collected literature.

2.1 Process

To find relevant publications, we used the search term *"nature inspired software"* on the catalogues of Google scholar, Elsevier, and Springer. We skimmed through the first 20 result pages and collected all publications in the biomimicry domain. Next, we recursively reviewed all citations and included all related cited works. Eventually, we collected 111 papers. Next, we classified our collected papers based on criteria inspired by the works of Kitchenham *et al.* [8]. For example, we have been interested in the publication date, the publisher, and the originating country.

2.2 Observations

We elaborate on metadata such as the publication date, the submission target, the publisher, the origins, and the contribution.

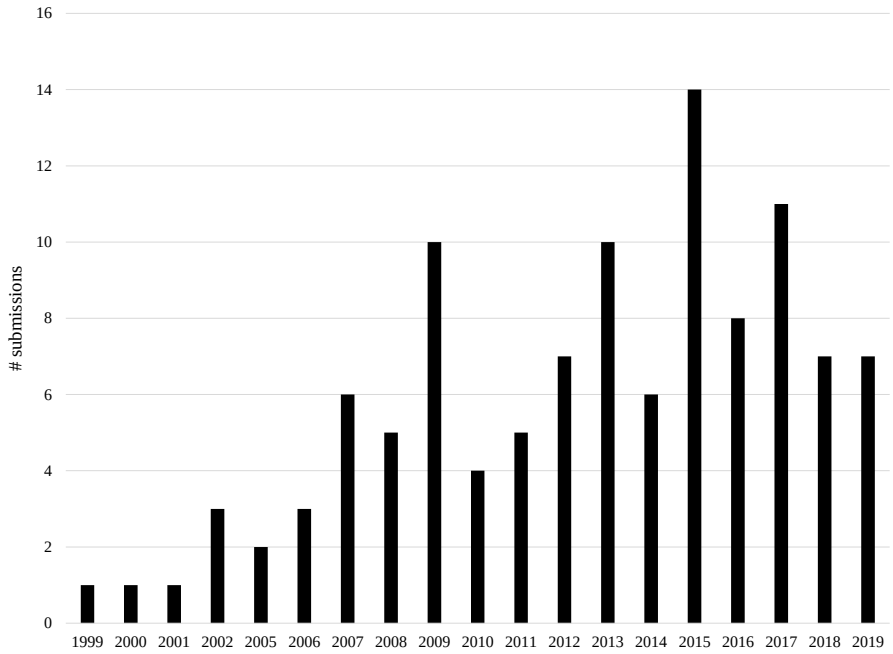


Figure 2.1: The interest in biomimicry research over time

2.2.1 Publication Date

We are interested in how the interest evolved over time to see whether this research area is still an emerging topic. In Figure 2.1, we present the year of release for all publications in our dataset. The y-axis denotes the number of publications we found. Our dataset contains publications between 1999 and 2019, and we can clearly see that the interest has increased during the past 20 years with a peak in 2015. The 14 publications of 2015 were submitted from 9 different countries, and most of them were either proposing an original concept (5), improving an existing proposal (4), or adapted an existing algorithm for a specific use case (4).

2.2.2 Submission Target

Most publications in our dataset underwent a strict review: 68 papers (61%) targeted a journal, whereas only 43 papers (39%) targeted a workshop or conference.

2.2.3 Publisher

The papers have been published by 20 different publishers, namely *Elsevier* (59, 53%), before IEEE (24, 22%), and Springer (9, 8%). The rest were published by local publishers from different countries (*e.g.*, India, Germany, and Helsinki), and online publishers such as *arXiv*.

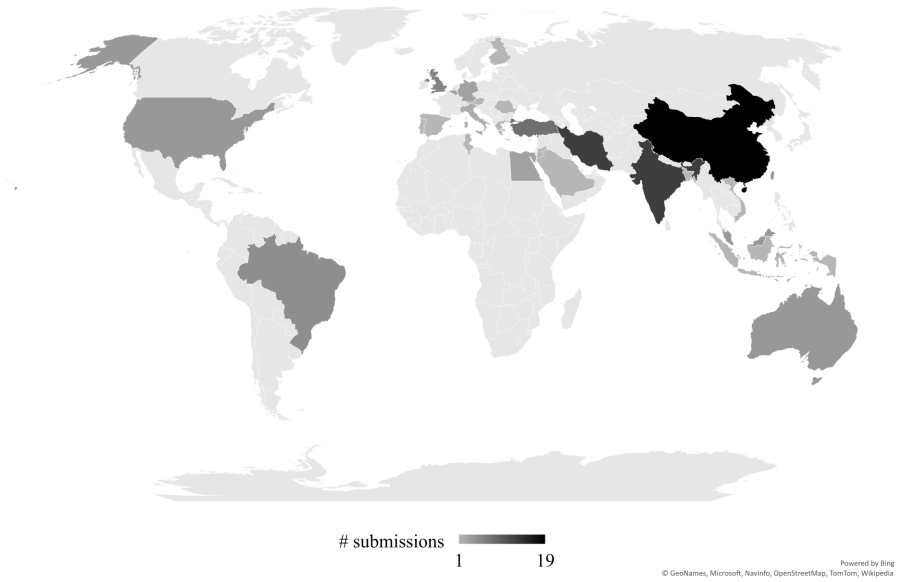


Figure 2.2: Participating countries in biomimicry research

2.2.4 Origins

We were interested in finding countries that host institutions which are specifically involved with biomimicry. Hence, we identified the corresponding country for each paper in our dataset by resolving the location of the first author's institution. In Figure 2.2, we present a world map with countries that participate in biomimicry research. The number of publications determines the color of the respective country in the map. Countries with no assigned publications are indicated with the lightest shade of grey. We can see that China is very active in this domain with 19 publications (17%), followed by India and Iran each with 13 publications (12%). Surprisingly, European countries are less prevalent. For example, Turkey is only responsible for 8 (7%), the United Kingdom for 6 (5%), and Germany and Belgium are each accountable for a mere of 3 publications (3%).

2.2.5 Contribution

In order to select relevant publications for further evaluation, we determined the type of their contribution, *i.e.*, a novel algorithm, concept, framework, or a comparison, survey, or review. The majority of research focuses on algorithms (80%), and much less on reviews (6%), concepts (5%), frameworks (4%) and surveys (4%). Only few works presented a comparison between two different algorithms (2%). The lack of diversity for comparisons might pose a threat to validity for such research. Furthermore, we classified the algorithmic contributions into the three subcategories *new*, *improvement*, and *potential use*. We found 51 publications that improve an existing algorithm, *e.g.*, tweaking parameter configurations or

integrating additional criteria, 23 publications presented a potential use case, and only 15 publications proposed a novel algorithm. Most novel algorithms received at least one follow up publication dedicated to improvement. China's researchers were particularly interested in advancing existing algorithms: 15 publications (88% of all Chinese papers) targeted improvements of prior work.

3 Findings

In this section, we focus on the proposed algorithms, their underlying theories, and evaluations to answer **RQ**$_2$: *What are characteristics of the proposed algorithms in existing literature, and how comprehensive is their support?* We reveal white spots in the research landscape and eventually show the lack of generalization in algorithmic publications. Our criteria for the algorithms are inspired by the work of Izto *et al.* [5], however we collected more features and provide a more accurate analysis. For example, we classified the publications according to the different domains in nature, and we determined distinct features such as the species, habitat, and pack size.

3.1 Classification of Life

In Figure 3.1, we present the classification of life that inspired the design of the algorithms. The x-axis denotes the classification of the algorithms, whereas the y-axis indicates the number of corresponding papers. For seven publications we were unable to determine a category, *e.g.*, when a proposal was related to genetics. We observe that categories containing species with very large numbers of collaborating individuals, *e.g.*, *insects* and *birds* are preferred for modeling algorithms. *Mammals* live in communities with fewer individuals, however they show complex social behavior and coordination that has been of interest for many researchers, *e.g.*, the collective search for prey. The interest in *plants* was mostly about their reproduction through seeds. Rather primitive life forms such as *viruses* and *protista* only received

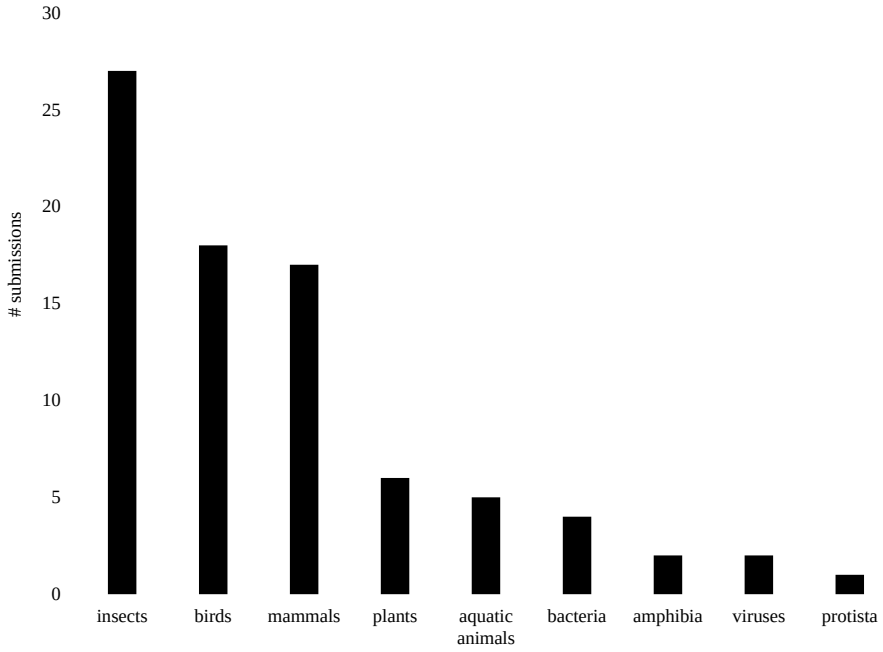


Figure 3.1: Prevalence of publications related to the domains in our dataset

little attention for their abilities to spread across an organism.

We see much potential in habitats that host intelligent species, but only received little attention. For example, we assume that the category *aquatic animals* provides many opportunities for future research.

3.2 Identified Species

In Figure 3.2, we present the particular species that inspired the design of the algorithms. The x-axis denotes the classification of the algorithms, whereas the y-axis indicates the number of corresponding papers. As before, we were unable to determine the value for seven publications. We can see that *bee* dominates as source of inspiration for biomimicry algorithms, closely followed by *bird*. The fact that bees are so outstandingly represented compared to the other species is surprising. According to Abbass [1], the reason is that they exhibit many features that distinguish their use as models for intelligent behavior, *i.e.*, division of labor, communication on the individual and group level, and cooperative behavior. The *bird (unspecified)* species provides various exploration and swarm behaviors that have been adapted to many different applications, *e.g.*, finite element modeling, or load scheduling in cloud computing. Bats received much interest in their echolocation abilities that have been adapted in many different forms, *e.g.*, the concept of echolocation combined with chaotic behavior to improve the search mobility for more robust global optimization. Cuckoos and grey wolves share a similar interest due to their complex social behaviors, *i.e.*, wolves have strict social rules when hunting for prey, whereas the cuckoos reported in the

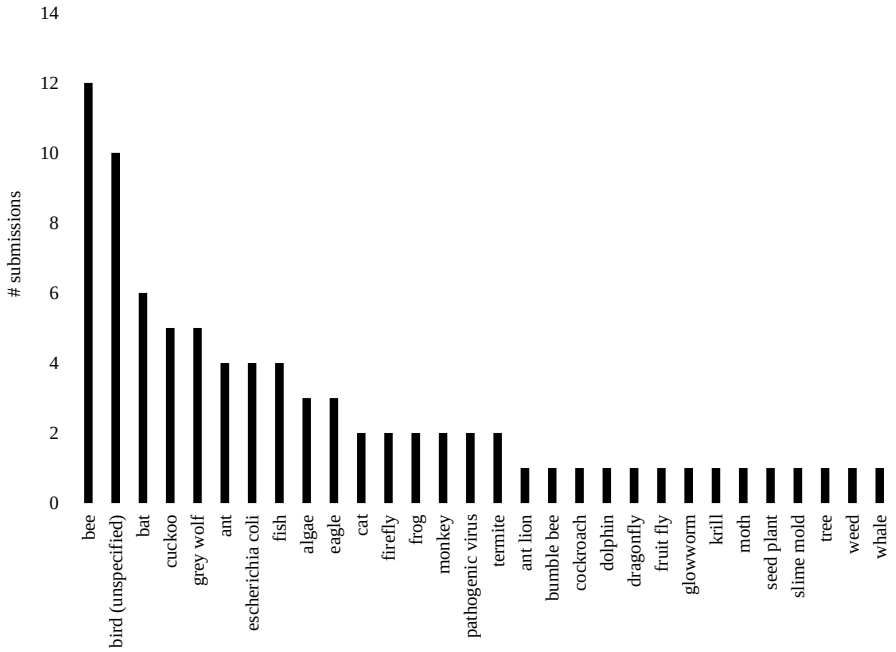


Figure 3.2: Interest in particular species

scientific work live their early life as brood parasites. At the end of the spectrum, the representation count of the species drops quickly and we find that twenty of our thirty identified species are only represented by one or two publications, *e.g.*, fruit fly, krill, or whale. All of these publications present a novel strategy, however there exists no literature that further improved them.

Since there exist many life forms, especially insects, whose behavior resembles that of bees we believe that there remain many opportunities for new discoveries in this field that could lead to improvements of the current state.

3.3 Habitat

We pragmatically categorized the problem domain of the proposed algorithms to better understand the problem scope. We could distinguish 12 categories including *generic optimization* which comprises solutions for abstract mathematical problems, *e.g.*, multi-modal numerical optimization. Next, we labelled the considered life form in each proposal with respect to its natural habitat, *i.e.*, *land*, *water*, *air*, or *intestines* and related it to the problem domain. We could not determine the habitats for eight papers, because that information was unavailable.

In Figure 3.3, we present the prevalence of habitats across different problem domains. The x-axis denotes the number of algorithmic proposals, whereas the y-axis lists the found types of problems. We can clearly see that *air* is the most common habitat, before *land*, *water*, and *intestines*. However, the habitat *air* has

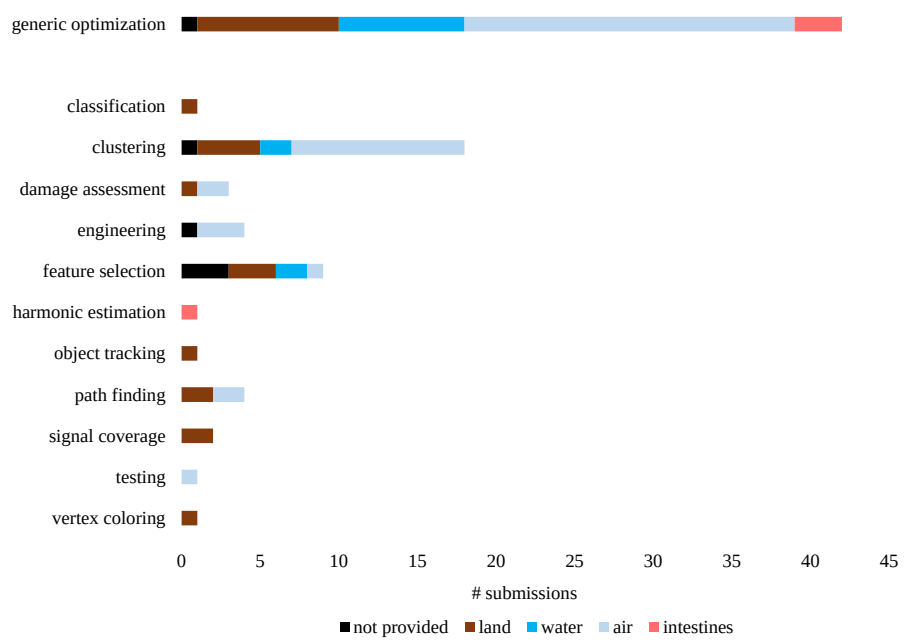


Figure 3.3: Preferred habitats for different problem domains

currently not been adapted to most problem domains as it only supports seven types of problems (58%). The most domains are supported by algorithms that emerged from the *land* habitat, *i.e.*, 9 of 12 problem types (75%) are supported. The most specialized habitat is *intestines* which only supports two problem types (16%).

Interesting for further research are problem types addressed in only a few papers and considering only a few different habitats. For example, we expect that problems in the domain of *classification*, *harmonic estimation*, *object tracking*, *testing*, and *vertex coloring* could greatly benefit from insights gained in other habitats. In particular, *generic optimization*, *clustering*, and *feature selection* received diverse interest and could provide valuable ideas for them.

3.4 Pack Size

We determined the pack size of the species for each paper, because we were interested in the researchers' use of scale. We first elaborate on the distribution of pack sizes for different problem domains, before we investigate the pack sizes across different habitats.

3.4.1 Distribution Across Problem Domains

In Figure 3.4, we show the prevalence of pack sizes for each problem domain. The x-axis presents the different problem domains, whereas the y-axis indicates the number of corresponding papers. Each bar

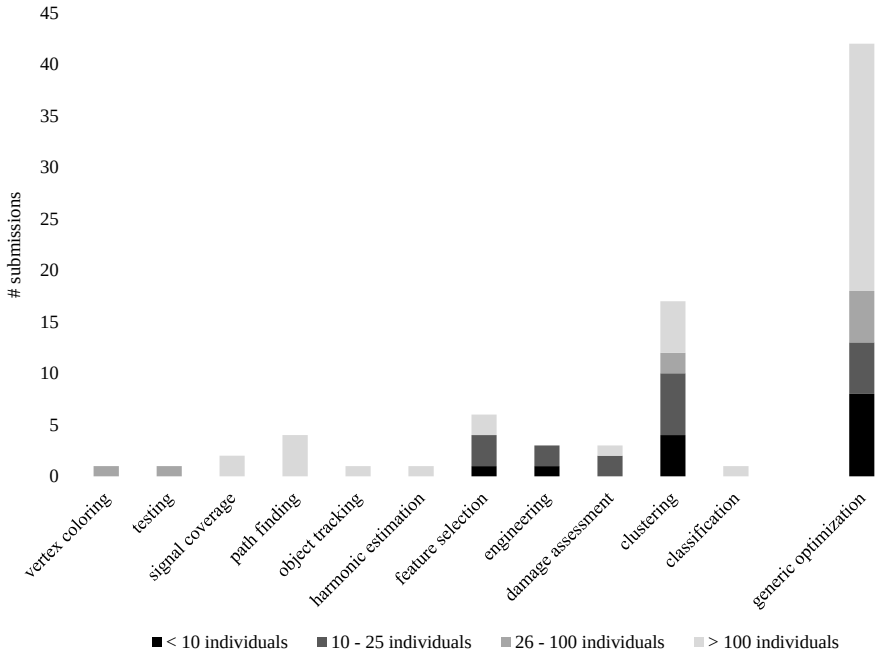


Figure 3.4: Representation of pack size in different fields

reveals the proportional use of the four different pack sizes: fewer than 10 individuals, 10 to 25 individuals, 26 to 100 individuals, or more than 100 individuals. We could not determine the habitats for seven papers, because that information was unavailable. The very large pack size is the most prevalent and suits the needs of nine different problem domains. All smaller pack sizes only support up to five different problem domains. Interesting is also their distribution: preferred are very large pack sizes with $>$ 100 individuals (50%). The remaining pack sizes received much less interest: the large, medium, and small pack sizes only gained 11%, 22%, and 17%, respectively.

Clustering, which encompasses algorithms for the organization of resources demonstrates less interest in large or very large pack sizes, *i.e.*, the small or medium pack sizes are the most prevalent. We found that this is due to increased interest in single but complex interactions between individuals, rather than the instinct-driven interactions of less intelligent individuals. Similarly, *feature selection*, *engineering*, and *damage assessment* prefer medium sized pack sizes. The solutions of these problem domains have in common that they define an initial suboptimal (local) solution and then actively begin to search for better solutions from that starting point. Such active searches require awareness that is only found in species with a higher intellect that usually live in smaller packs.

These findings reveal that the intellect of life seems to be an important factor to find solutions for rather distinct complex problems. However, further research is required to exclude any potential bias introduced by the researchers' assumptions.

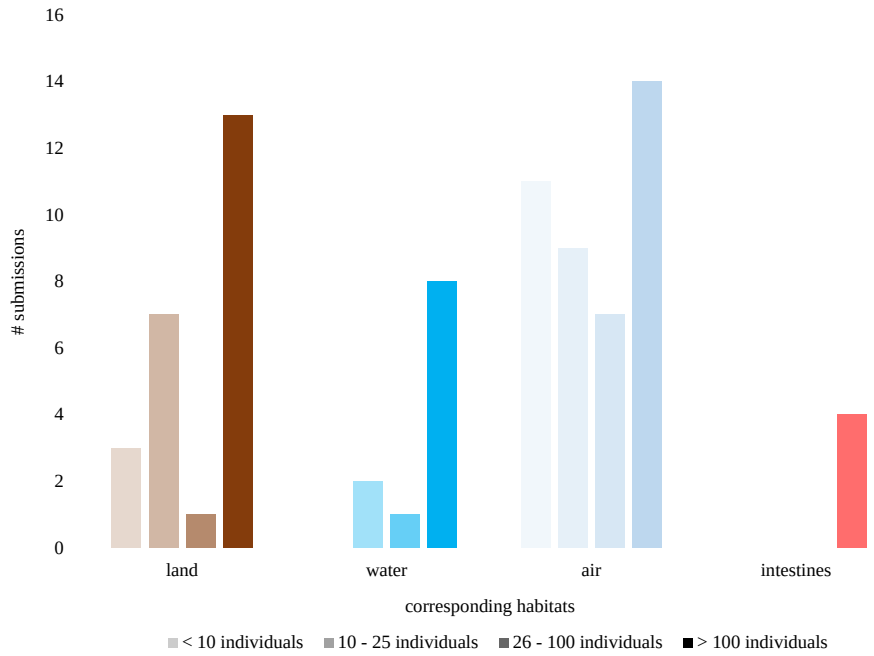


Figure 3.5: Prevalence of pack sizes for each habitat

3.4.2 Distribution Across Habitats

In Figure 3.5, we show the prevalence of pack sizes for each habitat. The x-axis denotes the different habitats and their use of pack sizes, whereas the y-axis indicates the number of papers. We could not determine the habitats for nine papers, because that information was unavailable or ambiguous. *Air* is the most dominant habitat, *i.e.*, it holds the most publications (51%) and it also claims the largest variety in terms of pack size. *Land* is the second most dominant habitat (30%) followed by water (14%) and intestines (5%). However, for *land* and *water* only few small and large pack sizes have been explored. *Intestines* primarily hosts countless microorganisms and thus we do not see any smaller pack sizes.

Water seems to be underrepresented and research provides further evidence for our discovery: terrestic life is better researched than aquatic life in which 91% of all species are expected to be unknown [10]. Moreover, we see a lack of interest for small and large pack sizes in *land*. Therefore, we encourage researchers to perform future studies on species from these domains.

3.5 Popularity

We are interested to see how long-lasting the interest of novel biomimicry concepts is. In Figure 3.6, we show the popularity of the three most prevalent species over time, *i.e.*, ants, birds, and bees. The x-axis denotes the time, whereas the y-axis indicates the number of papers. The first ant-related publication in our dataset is from the year 1999. The highest interest emerged three years later, and during subsequent

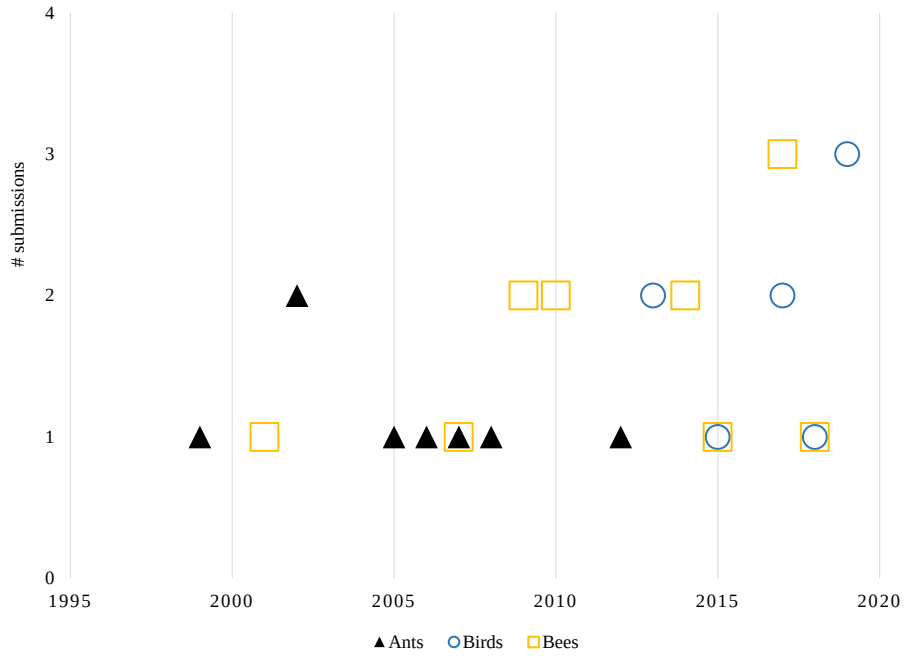


Figure 3.6: The popularity of the three most prevalent species over time

years follow up literature has been released, but at a lower pace until the year 2012. On the other hand, the interest in bees emerged in the year 2001, and has increased eight years later. Unlike the ant-related concepts, research of bees still remains a hot topic even 17 years later. Finally, the first biomimicry publication related to birds is from the year 2014; the interest has increased four years later. The research community remains very active and at least one publication has been released in every recent year.

The activity of certain biomimicry approaches seems to depend heavily on the flexibility of the applications and the quality of the result. Bees seem to provide many benefits, particularly more than ants. The data we have for birds is not enough to establish any long-term claims. We found that the interest increases three to eight years later after the initial publication. Based on the current data, we expect many bee and bird related publications in the near future, but only few related to ants, if any.

3.6 Efficiency Gains

We are interested in the achieved efficiency gains, because such knowledge could guide future prioritization of problem domains, *i.e.*, researchers could investigate domains with a lack of efficient solutions. For the efficiency classifications, if possible, we relied on the numerical values found in the proposals that compare existing work. If the numbers were not accessible, we searched in the text for efficiency-related statements from the authors. In Figure 3.7, we present the efficiency gains across different problem domains. The x-axis denotes the number of algorithmic proposals, whereas the y-axis enumerates the different domains.

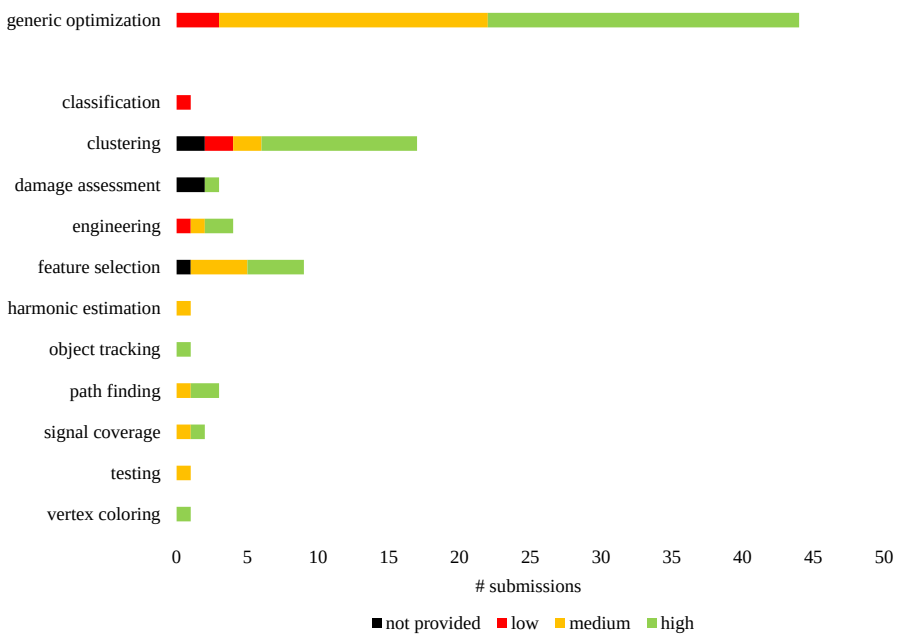


Figure 3.7: Reported efficiency gains for the different problem domains

Each bar consists of up to four different sections: black indicates a lack of information, red indicates a low gain (up to 5% better efficiency), yellow indicates a medium gain (up to 10% better efficiency), and green indicates a high gain (more than 10% better efficiency). The most claimed efficiency gain is *high* (52%), followed by medium (34%), and low (8%). We were unable to find any efficiency claims for five publications (6%). The distribution of the different levels of efficiency gain in each category is rather consistent: medium and high gains are for most problem types in majority. In other words, such gains have been achieved in 88% of all papers. However, *classification* did not benefit from such efficiency gains. We further found that 82% of the papers benchmarked their proposals against algorithms from generic optimization, 15% benchmarked only against their adapted original algorithm, and 3% did not provide any benchmarks.

We can see that *classification* lacks an efficient biomimicry-inspired solution. Moreover, researchers achieved in the problem domains *harmonic estimation* and *testing* only mediocre gains. Finally, almost every fifth publication did not perform a comprehensive evaluation.

3.7 Generalization

The ability to generalize existing work increases its value for the community, because such work may contribute to other problem domains. For each algorithmic publication, we skimmed through the title, the abstract, and the introduction to find information regarding the support for generalization, *i.e.*, whether

the researchers provide any information that a proposed algorithm solves a generic problem. That is, we searched for statements such as "performs the numerical optimization of multi-dimensional features." Whenever we found no indication of such support, we count that as support being absent. After reviewing all algorithmic publications, we found that 38% lack generalization in the title, 26% in the abstract, and 22% in the introduction. Moreover, 38% lack generalization in at least one section, and 20% in all three sections. Usually, the lack of generalization was present when very specialized statements existed, *e.g.*, "the purpose of this algorithm is to *optimize the load-balancing of cloud services*."

The lack of generalization hinders researchers from finding potential algorithms for their problems. Hence, if properly adapted, we expect numerous viable alternative techniques for a particular problem domain.

3.8 Replicability

In order to leverage a proposed algorithm, its implementation must be reproducible. For each proposal we investigated the existence of the provided implementation, *i.e.*, whether the source code exists, and its writing style. We found that 22% did not provide any pseudo code, and that 1% provided pseudo code in prose. 73% of all papers provided pseudo code that could be used for further adaptation.

More than every fifth paper did not provide any pseudo code. This makes it unnecessarily hard for practitioners to adapt a proposed solution.

3.9 Discussion

Biomimicry represents rather a small subset of algorithmic research. However, the merits are considerable in the field of metaheuristics for very complex problems that frequently occur in connected and distributed systems. In the proposals, we discovered a dominance of the habitat *air*, and a much lesser prevalence of the habitat *water* that is still left to be explored. For example, the *aquatic animals*, which are in our data primarily represented by fish, contain many species which express a swarming pattern similar to bees, ants, and birds. However, there are much fewer publications concerning them in comparison to insects and birds.

Similar strategies can be found in quite dissimilar species from different habitats. As a result, there is evidence that every problem domain can learn from species of different habitats. In particular, we expect that most proposals relying on the same pack size are interchangeable to a certain degree. Moreover, we expect additional gains when combining one or more strategies to leverage their advantages, while limiting the impact of the drawbacks a single strategy might entail. We envision a similar adaptation process as for algorithms related to bees: a single idea sparked seven publications across different problem domains.

We found a particular lack of information for many proposals. For example, gains have not been clearly reported, evaluations were not comprehensive, or generalization has not been any concern.

It is important that existing literature is properly found, understood, and used, which currently seems to

not be the case. That would require that generalization be considered an important factor for future work. We exemplify the missed potential of a very specific algorithm for load scheduling by creating a generic adaptation that could be used for purposes beyond load scheduling, and reveals its true purpose: clustering.

4

Adaptation of Existing Knowledge

In this section, we answer $\mathbf{RQ}_3$: *How can we generalize a specialized algorithm, and what are the gains?* We want to show that a highly specialized algorithm can be generalized to be used in other contexts. We implemented such an algorithm and applied it to a clustering problem. We finally show that the individual components of the algorithm indeed contribute to its success, even beyond the sole purpose of load scheduling.

The generalization itself is a manual process and requires basic mathematical knowledge in order to identify the actual purposes. Usually, related keywords are used that can help to identify them, *e.g.*, for clustering the words "grouping", "scheduling", *etc.* Papers that are suitable for generalization are usually highly specialized and propose an algorithm for a particular technical problem, *e.g.*, load scheduling, software testing, or mining data. As a result, we were interested in algorithms that provide pseudo code, but lack generalization in the title, the abstract, and the introduction. We randomly chose the algorithm reported in the paper "Cost optimized Hybrid Genetic-Gravitational Search Algorithm for Load scheduling in Cloud Computing" (HGGSA) by Divya *et al.* [2], which matches these criteria. The original publication proposes the algorithm exclusively for load scheduling in cloud computing, based on genetics and gravity. However, we realized that the algorithm itself is designed for clustering, but this information is unavailable in the publication. In fact, the entire publication does not even mention once the word "cluster." After the actual purpose has been identified, we started with the implementation.

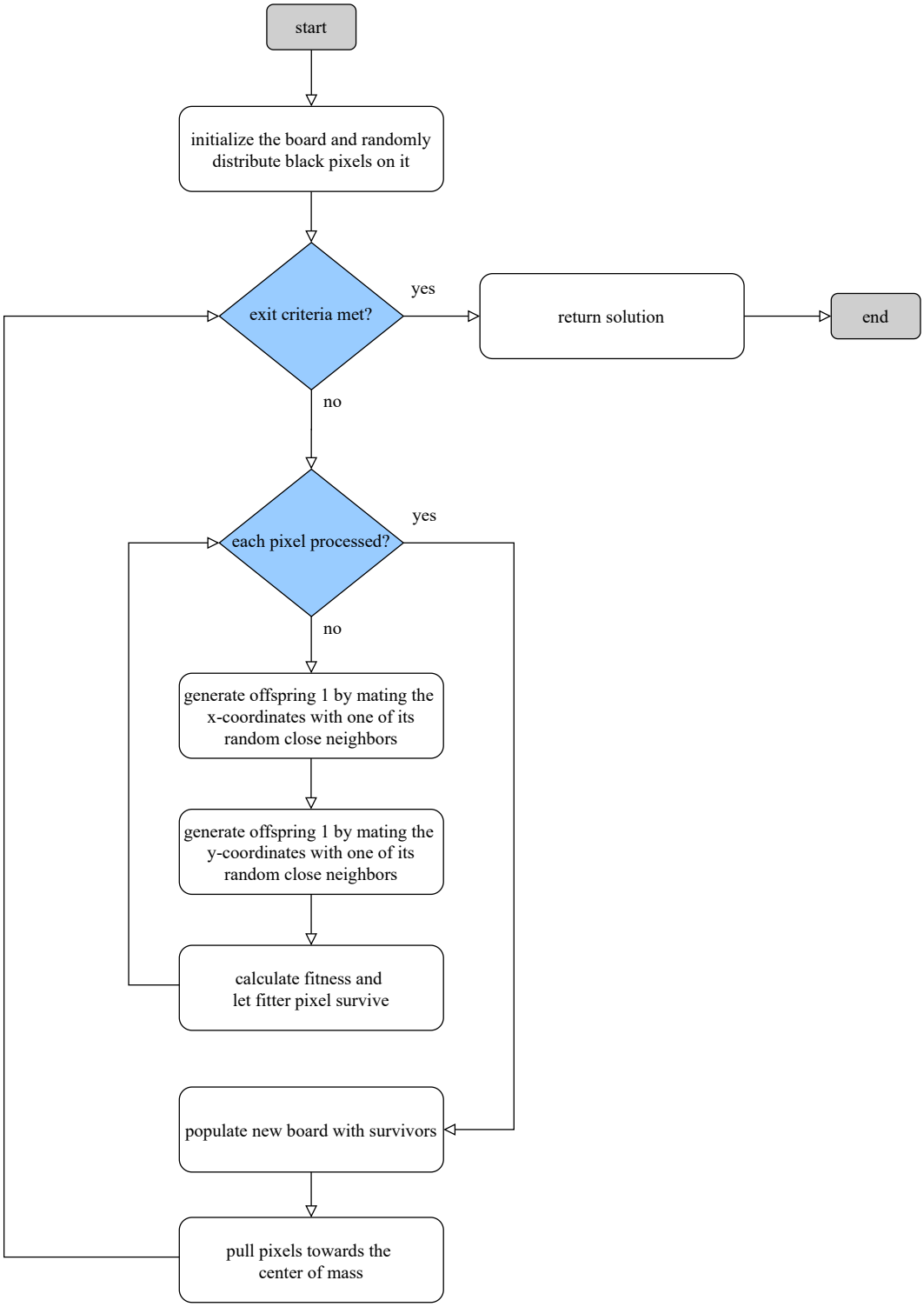


Figure 4.1: The concept behind the hybrid genetic-gravitational search algorithm

4.1 Implementation

We used the Java programming language for the implementation, and we closely followed the pseudo code presented in the original paper shown in Figure 4.1. We constructed a whiteboard view that can visualize each step of the optimization process. The whiteboard view relies on a two-dimensional boolean array that maintains the state and is initially populated with some random `true` values. Each value, `true` or `false`, represents a black or white pixel on the whiteboard, respectively. The *fitness* is calculated as how many other black pixels are adjacent to a specific pixel.

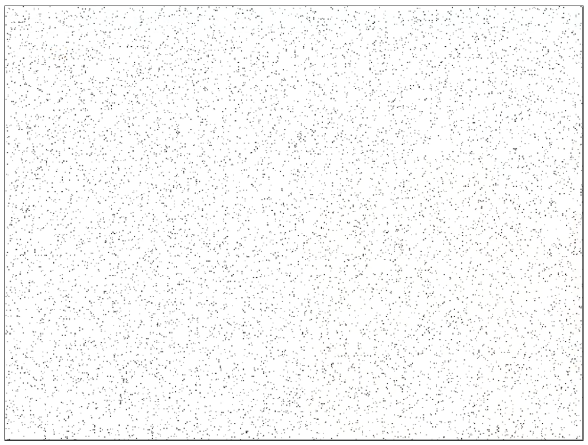
The proposed algorithm uses a combination of existing algorithms, *i.e.*, a genetic and a gravitational search algorithm. The genetic function first generates new pixel coordinates by mating a black pixel with another black pixel in its vicinity. This process creates two offspring pixels by crossing the x and y coordinates of the parents, and of those two pixels only the child with the higher fitness survives and does not turn white. Consequently, for the next iteration the parents are eradicated as well. Next, a gravitational force is applied that pulls the survivors one field towards the center of mass to accelerate the process of clustering. The current state is returned, if the reduction of the population reaches a predefined threshold. Otherwise, the procedure iterates one more time with the survivor pixels as new parents until the end criteria is met.

A typical run of our implementation is shown in Figure 4.2. At first, 20 000 black pixels are placed randomly on the whiteboard as presented in 4.2a. Next, the reduction of black pixels to clusters starts. 4.2b shows the process midway in which the majority of black pixels already has been assigned to a particular cluster. Finally, after 205 iterations at the end of the execution we can see in 4.2c that the black pixels are reduced to few clusters in the center of mass.

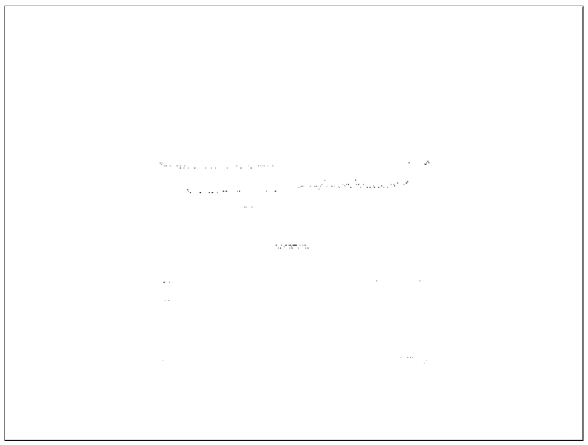
4.2 Evaluation

We repeated the measurements ten times each time using a different randomly generated board, and calculated the mean value. We used the following configuration: 800 by 600 pixels board size, 20 000 random black pixels as starting condition, gravity constant set to two, and the coordinates of the center of mass set to (400,300). In order to make the individual contribution of each of the combined algorithms tangible, we ran the same experiment three times: once using the genetic algorithm (GA), once using the gravity search algorithm (GSA), and once using a combination of both (HGGSA).

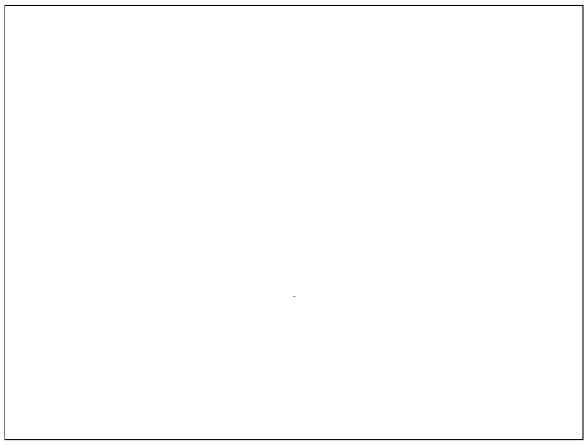
Figure 4.3 presents the results. The x-axis denotes the number of iterations, whereas the y-axis presents the number of separated clusters present at a particular iteration. The lines represent the three algorithms. In general, we can see that every algorithm is able to cluster data and therefore reduces the number of black pixels over iterations. However, there exist substantial differences between the three algorithms. GA and HGGSA reduce the cluster count from 20 000 to 1 000 in only 16 iterations or less, compared to 195 iterations required for GSA. Whereas HGGSA and GA achieve an exponential efficiency, GSA only achieves a polynomial efficiency. HGGSA reached the optimum after 205 iterations, GSA after 403 iterations, and GA started to stagnate from the 32. iteration. This is an intrinsic property of GA that is



(a) initial configuration (0 iterations)



(b) progressed configuration (102 iterations)



(c) final configuration, a single cluster left at the center of mass (205 iterations)

Figure 4.2: HGGSA clustering progress

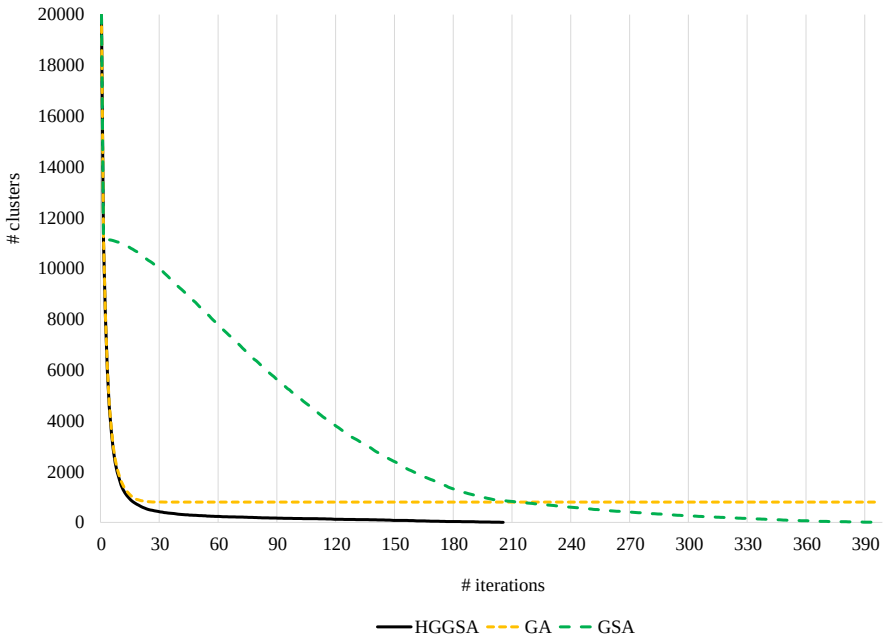


Figure 4.3: Performance of the generalized HGGSA

by design unable to converge, because it will build populations that are separate from each other, and since these populations eventually remain too far apart they cannot mate and thus result in non-converged clusters. Moreover, the time it takes to reach a converged state for the gravity-based algorithms depends on the used gravity constant; the higher it is set, the faster the system will converge. However, in the worst case, an inappropriate gravity setting prevents a converged state due to underfitting. In other words, “pixels would continuously move around the center of mass.”

In conclusion, HGGSA is able to maintain the initial momentum and quickly achieves a converged state compared to the other algorithms. Its genetic part reduces the distance between the local clusters and is able to connect different populations. On the other hand, the gravity part ensures that distant clusters collide over time, *i.e.*, the distance between pixels is continuously reduced.

5

Threats to Validity

An important threat to validity is the completeness of this study, *i.e.*, whether we could find and study all related literature. Although we could not review all papers, we aimed to explore top-tier biomimicry and applied computation journals and conferences as well as highly-cited work in the field. Moreover, we might have missed relevant criteria. We mitigated that threat by peer-reviewing the criteria catalogue.

Another major threat represents the correctness of the collected data, *i.e.*, whether the labelled data are accurate. We established criteria for the correct labelling of each property to prevent any ambiguities. Whenever a labelling task was not clear, the problem has been resolved by discussion with a supervisor.

Finally, the fact that the adapted algorithm is validated by the author is a threat to construct validity through potential bias in experimenter expectancy. We mitigated this threat by including a supervisor in the process.

6

Related Work

The concept of biomimicry is not new and some researchers already investigated the rationale behind seeking inspiration from nature. For example, Steer *et al.* reasoned about the features which are a valuable source for the design of successful nature-inspired algorithms [12]. This work complements our study: they elaborate on the process to find relevant algorithms, where we report what life has not received much attention and ask for a more generalized argumentation.

There also exist multiple books that cover nature-inspired algorithms [3, 13, 14]. Unfortunately, all of them are structured around the origins of the algorithms instead of their supported optimization problems. Such information is not much of use for practitioners. Furthermore, they do not provide an exhaustive list of published work, but instead they present a few selected algorithms and present them often in great detail.

Fister *et al.* performed a survey of nature-inspired algorithms [5]. They compiled a list of 78 algorithms for optimization. Contrary to our work, they focused primarily on the rudimentary classification of the different algorithms and barely collected features from them. Although they established a brief list of algorithms for practitioners, they do not provide any usable information on where to use them. We argue that generalization is key to find relevant algorithms for specific problems.

In summary, existing work provides only shallow information for practitioners and is not aware of the generalization problem. Information about generalization would ease the search for algorithms that suit a

particular optimization problem.

7 Conclusion

We have performed a literature review on 111 publications related to biomimicry and extracted several characteristics, *e.g.*, meta-data, the solution, and the investigated species. Our results indicate a structural issue: publications related to new or adapted algorithms very prominently emphasize on a specific use case, instead of the generalized problem category, *e.g.*, clustering. We found that 38% lack generalization in at least one of the introductory elements (*i.e.*, title, abstract, and introduction), and that 53% of them lack generalization entirely. Moreover, 40% of the proposed algorithms lack at least one major characteristic, *e.g.*, code samples or benchmarks against state of the art algorithms. We motivate the found generalization problem with our adapted implementation of an algorithm proposed for load scheduling. Moreover, the artifacts of this study can support practitioners in finding more efficient existing solutions across research domains.

Bibliography

[1] H. A. Abbass. MBO: Marriage in honey bees optimization - a haplometrosis polygynous swarming approach. In *Proceedings of the 2001 congress on evolutionary computation (IEEE Cat. No. 01TH8546)*, volume 1, pages 207–214. IEEE, 2001.

[2] D. Chaudhary. Cost optimized hybrid genetic-gravitational search algorithm for load scheduling in cloud computing. *Applied Soft Computing Journal*, 83(105627), 2019.

[3] R. Chiong. *Nature-inspired algorithms for optimisation*, volume 193. Springer, 2009.

[4] L. Ding, T. Finin, A. Joshi, R. Pan, R. S. Cost, Y. Peng, P. Reddivari, V. Doshi, and J. Sachs. Swoogle: A search and metadata engine for the semantic web. In *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management*, CIKM '04, page 652–659, New York, NY, USA, 2004. Association for Computing Machinery.

[5] I. Fister Jr. A brief review of nature-inspired algorithms for optimization. *Elektrotehniski vestnik*, 2013.

[6] R. Guha, R. McCool, and E. Miller. Semantic search. In *Proceedings of the 12th international conference on World Wide Web*, pages 700–709, 2003.

[7] A. H. Hofmann. *Writing in the biological sciences: A comprehensive resource for scientific communication*. Oxford University Press, 2013.

[8] B. Kitchenham, O. P. Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman. Systematic literature reviews in software engineering–a systematic literature review. *Information and software technology*, 51(1):7–15, 2009.

[9] Y. Lei, V. Uren, and E. Motta. Semsearch: A search engine for the semantic web. In *International conference on knowledge engineering and knowledge management*, pages 238–245. Springer, 2006.

[10] C. Mora, D. P. Tittensor, S. Adl, A. G. Simpson, and B. Worm. How many species are there on Earth and in the ocean? *PLoS Biol*, 9(8):e1001127, 2011.

[11] J. Schimel. *Writing science: how to write papers that get cited and proposals that get funded*. OUP USA, 2012.

[12] K. C. Steer, A. Wirth, and S. K. Halgamuge. The rationale behind seeking inspiration from nature. In *Nature-inspired algorithms for optimisation*, pages 51–76. Springer, 2009.

[13] X.-S. Yang. *Nature-inspired metaheuristic algorithms*. Luniver press, 2010.

[14] X.-S. Yang. *Nature-inspired optimization algorithms*. Academic Press, 2020.

A

Anleitung zum wissenschaftlichen Arbeiten

The Anleitung consists of the conference paper “Biomimicry-based Algorithms and Their Lack of Generalization”.[1]

P. Gadient, D. Klopsch, M. Ghafari, and O. Nierstrasz. Biomimicry-based Algorithms and Their Lack of Generalization.

Scheduled for submission to *International Conference on Metaheuristics and Nature Inspired Computing: META 2022*, 2022.

[1] http://scg.unibe.ch/download/supplements/Biomimicry-based-Algorithms-and-Their-Lack-of-Generalization-(Submission).pdf