Bee Nest-Site Selection Clustering

Aurel Paulovič*
Institute of Informatics and Software Engineering
Faculty of Informatics and Information Technologies
Slovak University of Technology in Bratislava
Ilkovičova 3, 842 16 Bratislava, Slovakia
aurel.paulovic@gmail.com

Abstract

This paper introduces the design of a novel data clustering algorithm based on the inspiration by the process of nest-site selection of honey bee colonies also known as swarming. Using the analogy between n-dimensional data space and country-side where the data points represent flowers, we can understand clusters as regions with great amount of food. Using the swarming analogy, the algorithm than creates the majority of abstract bee hives in areas with sufficient food resources, while the areas with food shortage become less populated, thanks to natural selection and resources race between the colonies. Interpreting the hives distribution we can than determine positions of respective clusters.

Categories and Subject Descriptors

I.2.8 [**Artificial Intelligence**]: Problem Solving, Control Methods and Search—*heuristic methods*; I.5.3 [**Pattern Recognition**]: Clustering—*algorithms*

Keywords

Data mining, clustering, swarm intelligence, swarming, bee hive, nest-site selection

1. Introduction

Data mining and knowledge discovery is a vivid field of study that receives more and more attention especially with the advent of large informations systems, web and big data. One of popular data mining techniques is clustering, which is an unsupervised learning technique for grouping data points into groups of similar items, or more formally, into groups with low intra-variance and high extra-variance.

During the last years, nature inspired algorithms have become increasingly popular and were used to solve many different algorithmic problems. One class of the nature algorithms, the swarm intelligence, that is characterized by its self-organization and high parallelization, appears to be especially suitable for solving various tasks of data mining.

In this paper we propose a novel data clustering algorithm inspired by the process of nest-site selection of honey bees in nature. In order to identify clusters, the algorithm uses the similarity between areas with high data density and country-side regions with rich food reserves to distribute abstract bee hives in data space and the nature selection and resources competition to gain self-repairability and lower the computational complexity.

The rest of the paper is organised as follows. Section 2 briefly reviews different clustering methods inspired by swarm intelligence with focus on honey bee algorithms. In section 3 we describe the process of nest-site selection of honey bees in nature. We present the design of our new nest-site selection clustering algorithm in section 4. Finally, section 5 concludes the paper and gives information about our further work on the algorithm.

2. Swarm intelligence in clustering

Many clustering methods have been developed and they can be broadly classified into 7 categories [2]: hierarchical, partitioning relocation clustering, density-based partitioning, grid-based methods, constraint-based clustering, neural networks and evolutionary methods. With the advent of swarm intelligence, we can also add the category of nature inspired algorithms.

Swarm clustering uses the inspiration by collective intelligence of social insects (but there are also some non-insect swarm algorithms, e.g. PSO [1]) to solve the NP-hard problem of finding clusters in large data. Some of the swarm inspired clustering methods use analogy with ants and their sorting behaviour [4], other use foraging or mating behaviour of bees [8, 3]. A detailed survey of honey bee algorithms in general can be found in [5].

*Master study programme in field Information Systems. Supervisor: Professor Pavol Návrat , Institute of Informatics and Software Engineering, Faculty of Informatics and Information Technologies, STU in Bratislava.
Work described in this paper was presented at the 7th Student Research Conference in Informatics and Information Technologies IIT.SRC 2011 and awarded by IEEE Czechoslovakia Section Award.

Paulovič, A. Bee Nest-Site Selection Clustering. Information Sciences and Technologies Bulletin of the ACM Slovakia, Special Section on Student Research in Informatics and Information Technologies, Vol. 3, No. 2 (2011) 100-103

Some of the bees inspired algorithms use the analogy of bees foraging behaviour. The algorithm proposed in [8] exploits the search capability of bees to overcome the local optimum problem of the general k-means algorithm. The honey-bee mating optimization algorithm [3] combines simulated annealing with the mating flight of a bee queen to find the best partitioning of a data-set. Flower pollination by artificial bees [6] is an extension of AntClass algorithm and uses the idea, that each species of plants has better growth in special region and agglomeration of these species in such places is observable. Using the pollination by bees, the algorithm copies the data to various places and then uses natural selection to select best agglomerations of the data.

3. Nest site selection of bees in nature

The life of a honey bee colony in cold regions takes place in annual cycles. According to [9] shortly after the winter solstice the colony begins to preare for the next year and starts to rear brood. At the late spring the colony starts to reproduce in that it produces drones and rears several new queens. Just before the first new queens are mature, the old queen leaves the hive with about half of the worker bees in a process called swarming. After a short flight, the swarm sits down on a tree branch and sends out scouting bees, whose job is to find suitable places for a new hive.

Each scouting bee flies alone and tries to find suitable place for hive establishment and if it finds such a place, it evaluates it briefly and returns back to the swarm. In selecting the right place scouting bees consider many different factors like its distance from the parent hive, which should be not too long, nor too short, proportions of the hive candidate and its temperature stability, etc. After his return, the scout informs the rest of the swarm about the position and estimated quality of the candidate hive position using a waggle dance, which is similar to the dance the bees use to communicate the locality of food sources when foraging [7]. Typically scouts find more than one candidate position and the dance serves to convince the other bees, that the promoted place is the best one, whereby the scouts dance for a candidate position more intense, if they think it is of a higher quality, and less, if it is not so good. Step by step the bees propagating worse places let them self get convinced and they abandon inferior candidate positions and start to dance for the better ones. At the same time, the more bees dance for some place, the more bees go to evaluate it themselves, which in turn lets them rate it more accurately and discover potential defects which may not be known at first (e.g. the place becomes infested by ants or if fills up with water after a rain). Usually the swarm eventually settles on a single candidate position and it will start a new hive there.

For a week or so following the departure of the old queen with a portion of the worker bees the rest of the colony is queenless, after that the first new virgin queen emerges [9]. If the colony is strong enough, the remaining workers will force the queen to leave in an afterswarm, otherwise they will let the queen kill all remaining queens that are still in their cells. Depending on the strength of the colony, it can support multiple swarmings till there is only one queen in the hive. In case that there are multiple queens and there are not enough worker bees to do more swarmings, workers will let the remaining queens fight each other until just one survives. After the mating flight of all the surviving virgin queens the reproduction phase of the annual cycle ends.

The swarming process of honey bees can be summarized in following points:

1. Old queen with a part of the workers leaves the old colony in a swarm.
2. The swarm sends scouts to find candidate positions for a new hive and rate them.
3. Scouting bees dance for respective candidate positions and superior places get more support, while inferior places loose support.
4. Superior places are more popular and are evaluated more thoroughly in an iterative fashion.
5. Inferior places are abandoned and the swarm eventually settles on a single candidate position, where it sets up a new hive.
6. If the originating colony is strong enough, it may support subsequent swarmings.

4. Bee nest-site selection clustering model

By the examination of the life-cycle and reproduction of a honey bee colony, it is possible to draw a parallel between the nest-site selection of bees and searching for clusters in data. We can say that bees search for a candidate hive positions in a 3-dimensional space, country-side, in which they try to find the best positions, by which they consider quality measures like distance from the parent and other hives and the amount of food sources. They try to find a place with minimal competition and maximal food resources.

We can understand the features (attributes, columns) of a dataset as individual axis of a n-dimensional data space, where each single element (feature vector) represents a point in this space, whose position is determined by the values of its features. For the sake of simplicity we will further consider only continuous numerical attributes. Elements of the dataset in n-dimensional space have a certain distribution structure that corresponds to the character and properties of the events that produced it. Therefore finding clusters in such data means for us the disclosure of this often hidden or non-obvious structure. In data space we will call clusters those regions, that show relatively high data density and that are separated from other clusters by a region with relatively low data density.

In nature meadows and orchards could be considered to be places with relatively large amount of food resources having high density of flowers, while places like desert or sea would represent regions with little or no food at all. Using the parallel with data distribution, if we take data elements as food resources, e.g. flowers, we can consider meadows and orchards to be clusters with high flower density separated from other such clusters by regions with shortage on flowers. Since bees will tend to build hives in regions with sufficient food reserves, we could say that the majority of hives would be situated in places with higher density of flowers, i.e. in clusters. On this basis we can say, that while not knowing the structure of clusters in advance and having a certain distribution of bee hives, we can infer the respective clusters from the positions and density of hives, where each cluster will be

denoted by a region with above average density of hives. Therefore by designing an algorithm that can create and insert new abstract bee hives in the data space we will create a clustering algorithm.

In order to be able to define our clustering algorithm we introduce following terms:

Hive area The area defined by the hive position and the radius in which all the flowers (data points) contribute to the vitality of the hive. As the hive position, i.e. the center of the hive area, we will use only existing elements of the data set, not an arbitrary point in the space. The radius should be rather smaller because we want the clusters to be regions with high density of hives, that means not a single hive for a single cluster.

Hive vitality The fitness of a hive determined by the number of flowers in its area and the position of other hives and their areas (flowers that fall into multiple hive areas contribute to the vitality of each hive by a fraction).

Swarming area The area defined by the swarming radius with center in the parent hive that represents the region which is searched by scouting bees for candidate hive positions.

The proposed algorithm, as stated in Algorithm 1, consists of 4 main phases: protohives initialization, swarming, natural selection and cluster evaluation. We will discuss each of the phases in more depth.

Algorithm 1 Bee nest-site selection clustering – skeleton.

```
1: U ← initialize protohives{protohives initialization}
2: repeat
3:   for all hive in U do {swarming}
4:     K ← find candidate positions for a new hive
5:     while number of candidate positions > 1 do
6:       r ← rate candidate positions and choose the
         worst
7:       Remove(K, r)
8:     end while
9:     Insert(U, K)
10:  end for
11:  if natural selection condition then {natural selec-
     tion}
12:    R ← choose hives to be eliminated by natural
       selection
13:    Remove(U, R)
14:  end if
15: until terminal condition
16: evaluate clusters from the hives distribution{cluster
    evaluation}
```

4.1 Protohives initialization

In the initialization phase of the algorithm we need to insert several hives in the data space from which we will start swarming in the next phase. There are several possible solutions to the initial hive distribution, e.g. trivial or systematic, each having its pros and cons. Trivial distribution is undemanding to computing resources but can lead to relatively slow start of the algorithm, different heuristics can lead to more optimal initial distribution, but are usually much more resources intensive and can be relatively tricky to design.

Depending on the data distribution, initial hive distribution and the swarming radius, we can get into state, where some data points or even clusters are not reachable by swarming because they are separated from other clusters by distance greater than the swarming radius. This problem can be eliminated either by more complex initialization, in that we would insert at least a single hive in each such unreachable region, or more simply, by a mechanism we call honey bee drift, which we introduce in the swarming phase.

4.2 Swarming

At the beginning of each swarming phase an abstract bee swarm flies out of each hive and sends scouting bees to search for candidate hive positions. In our algorithm we do not need to model individual scouts or the abstract swarm itself, instead we will just pseudo-randomly pick up k data points in the swarming area (with center in the parent hive), where k is constant or is determined by the vitality of parent hive. Candidate positions are picked up pseudo-randomly depending on the distance from parent hive and/or the number of hive areas intersecting the swarming area.

The candidate selection and the value of k allow us to tune the algorithm to perform an exhaustive or rather a faster greedy search. Using bigger values of k will let us compare more candidate positions at the cost of slower computation and the risk of getting stuck in very dense data areas, whereby lower k values lead to faster but shallow search. Since flowers, that belong to multiple hive areas contribute to the vitality of each of the hives only by a fraction, we can use this information to spread the hives more evenly in that we will prefer for candidate positions those flowers, that belong to minimum hive areas. We can understand it as resource competition between hives.

Selected candidate positions are evaluated in iterative fashion. We first rate each of the positions simply by the number of flowers in its area or by the hypothetical vitality of the possible hive using small radius that is only a fraction of the full hive area radius. We choose and discard the worst rated candidate position. Than we rate each of the rest candidates again using a slightly bigger radius and discard the worst one. We perform this iterative selection until we get a single candidate position which then becomes the new hive. The process is illustrated in Figure 1.

The swarming radius should be greater than the hive area radius but can not be unlimited, because it would defeat the locality of candidate position selection and would result in concentration of almost all the hives in fewer small areas with very high data density, which in turn would make the algorithm ignore less distinct clusters. However, limited swarming radius can lead to unreachable regions in data especially if we combine it with naive initial protohives distribution. In order to overcome this limitation, we propose the honey bee drift mechanism. The drift can occur on every single swarming with certain (low) probability and it means that the scouts will not choose candidate positions in the swarming radius, but they will choose candidate position in the whole data space instead, whereby they will focus on dataset elements that are not

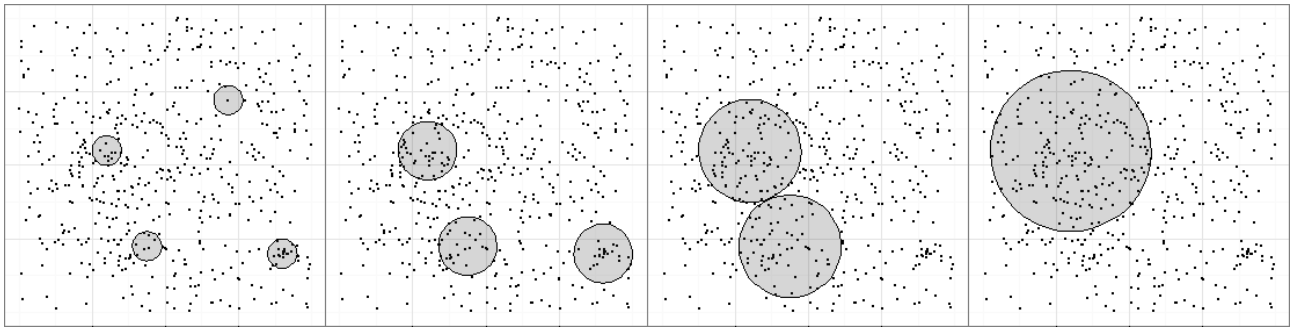

Figure 1: Iterative selection of best candidate position with increasing evaluation radius.

members of any hive area. This will allow us to create hives in otherwise unreachable areas and will not dramatically affect the locality of usual swarming.

4.3 Natural selection

The phase of natural selection serves as a regulation of needed computing resources and as a self-repair mechanism, otherwise if we would have swarming without natural selection, the number of hives and computation costs would grow geometrically. The goal of natural selection is to reach equilibrium, when the number and position of hives will be more or less stable, which will also be the terminal condition. In order to achieve a stable state we need to balance the production of new hives via swarming with their destruction by natural selection.

In this phase we can use different systematic or stochastic approaches. One of them is to simply remove all the hives, that have lower than some given minimum vitality, which means that we will remove hives in areas with very low data density or hives in overcrowded regions due to resources competition.

We can also arbitrarily destroy a group of vital hives and let the algorithm model repair itself. If the region with destroyed hives had high data density, the algorithm will spawn new hives quickly in this area because of no or only minimal resources competition and sufficient food resources, otherwise the region did not belong to some cluster, so we do not need to build hives there.

4.4 Cluster evaluation

The algorithm will build more or less coherent groups of hives in the data space denoting the positions and borders of data clusters. Clear clusters with higher data density will have more hives and sharper contours, less distinctive clusters will have less hives with greater mutual distance and regions without clusters will have only scattered hives with very low vitality. From the final hive distribution we can infer the resulting clusters from the superposition of hives in respective groups or by taking the marginal hives of individual groups as the surface of cluster body in n-dimensional space.

5. Conclusion and future research

This paper has presented a new clustering method based on the nest-site selection process of honey bee. The algorithm consists of 4 main phases: the protohives initialization, swarming, natural selection and the cluster evaluation. The proposed model does not require a priori knowledge of the number of clusters, allows for arbitrary shaped clusters and provides means to overcome locally optimal solutions.

In future research we plan to investigate the model further and build a prototype, that will implement the algorithm. Subsequently we will evaluate the model performance on various data-sets and compare it with other swarm clustering algorithms as well as with other well-known reference algorithms.

Acknowledgements. This work was partially supported by the Slovak Research and Development Agency under the contract No. APVV-0208-10.

References

[1] A. Abraham, S. Das, and S. Roy. Swarm intelligence algorithms for data clustering. In O. Maimon and L. Rokach, editors, *Soft Computing for Knowledge Discovery and Data Mining*, pages 279–313. Springer, 2008.

[2] P. Berkhin. Survey of clustering data mining techniques. Technical report, Accrue Software, San Jose, CA, 2002.

[3] M. Fathian, B. Amiri, and A. Maroosi. Application of honey-bee mating optimization algorithm on clustering. *Applied Mathematics and Computation*, 190(2):1502–1513, 2007.

[4] J. Handl and B. Meyer. Ant-based and swarm-based clustering. *Swarm Intelligence*, 1(2):95–113, 2007.

[5] D. Karaboga and B. Akay. A survey: algorithms simulating bee swarm intelligence. *Artif. Intell. Rev.*, 31(1-4):61–85, 2009.

[6] M. Kazemian, Y. Ramezani, C. Lucas, and B. Moshiri. Swarm clustering based on flowers pollination by artificial bees. In *Swarm Intelligence in Data Mining*, Studies in Computational Intelligence, pages 190–202. Springer-Verlag New York, Inc., 2006.

[7] M. Lindauer. Schwarmbienen auf wohnungssuche. *Journal of Comparative Physiology A: Neuroethology, Sensory, Neural, and Behavioral Physiology*, 37(4):263–324, 1955.

[8] D. T. Pham, S. Otri, A. Affify, M. Mahmuddin, and H. Al-Jabbouli. Data clustering using the bees algorithm. In *Proceedings of the 40th CIRP Int. Manufacturing Systems Seminar*, Liverpool, UK, 2007.

[9] T. D. Seeley. *The Wisdom of the Hive: The Social Physiology of Honey Bee Colonies*. Harvard University Press, 1996.