

# Appendix B - Annotated Bibliography

Joshua M. Tuggle\*

May 3, 2018

CS 6405 - Clustering Algorithms

\*Josh earned his Bachelor of Science in Computer Science in May 2016. He is a former member of the Natural Computation Laboratory at the Missouri University of Science and Technology. Josh hopes to graduate with a Master of Science in Computer Science in May 2018.

## CONTENTS

<b>I</b>	<b>Survey of Clustering Algorithms</b>	<b>3</b>
<b>II</b>	<b>Clustering</b>	<b>3</b>
<b>III</b>	<b>Introduction to Evolutionary Computing</b>	<b>3</b>
<b>IV</b>	<b>Improving Performance of SAT Solvers by Automated Design of Variable Selection Heuristics</b>	<b>3</b>
<b>V</b>	<b>A Comparison of Programming Languages in Economics</b>	<b>3</b>
<b>VI</b>	<b>Binaryheap: An implicit binary tree</b>	<b>3</b>
<b>VII</b>	<b>Introduction of Information Retrieval</b>	<b>4</b>
<b>VIII</b>	<b>UCI Machine Learning Repository</b>	<b>4</b>
<b>IX</b>	<b>KMeansGP</b>	<b>4</b>
	<b>References</b>	<b>5</b>

## I. SURVEY OF CLUSTERING ALGORITHMS

In [1], the authors provide quite a wide array of information on clustering algorithms. The thing I really used this source for was understanding the use of Euclidean distance. The thing that I cited in my paper was the fact that Euclidean distance is used with k-means clustering. Interestingly enough, I think it is fairly interesting that Euclidean distance is a special case of Minkowski distance.

## II. CLUSTERING

In [2], the authors provide a textbook on the topic of Clustering. Hence the name of the textbook. This source was used to investigate some of the time complexities involved with k-means clustering. I used this the information from this textbook to try and articulate the time complexity issue with the reader.

## III. INTRODUCTION TO EVOLUTIONARY COMPUTING

In [3], the authors provide fair sized book on the topic of Evolutionary Computing. I was taught in Dr. Tauritz's Evolutionary Computing course with this book. It taught me a lot of what I know today about EA/GP. My primary use for this project was to help me refresh my understanding of the fundamentals. Chapter 6 was most helpful, as it is an entire chapter on Genetic Programming.

## IV. IMPROVING PERFORMANCE OF SAT SOLVERS BY AUTOMATED DESIGN OF VARIABLE SELECTION HEURISTICS

In [4], the authors of this paper use GP to try and evolve heuristics for the minisat SAT solver. I worked on this project NC-LAB on the MST campus. After working on this paper/project, it was a significant help in working on this project.

## V. A COMPARISON OF PROGRAMMING LANGUAGES IN ECONOMICS

In [5], the authors show the speed of C++ versus other languages. I used this to show that C++ is a fast language. One thing that was interesting is that the compiler choice for C++ is fairly important for speed.

## VI. BINARYHEAP: AN IMPLICIT BINARY TREE

In [6], the author shows a way to represent a binary tree using an array. I tried this for ease of use of storing my trees in txt files and keeping track of children was fairly simple. It was painful to implement subtree crossover with the array representation. I'd say this method is easier to implement compared to trees with pointers.

## VII. INTRODUCTION OF INFORMATION RETRIEVAL

In [7], the authors briefly discuss the topic of purity. The purity way for checking the accuracy of my clustering solutions made sense and it was fairly easy to implement into my framework.

## VIII. UCI MACHINE LEARNING REPOSITORY

In [8], authors of datasets upload them to the UCI Machine Learning Repository for the world to use for various applications. I used this website to grab the Iris, Wine, and Isolet datasets. It is an easy to navigate site. I was sure to follow their citation policy and mentioned UCI in my acknowledgements.

## IX. KMEANSGP

In [9], the author goes over his head trying to write something from scratch instead of using premade tools and libraries. It was a fun experience writing the code for this project. I did not get the experimental results I wanted, but that is ok.

## REFERENCES

- [1] R. Xu and D. Wunsch, "Survey of clustering algorithms," in *IEEE Transactions on Neural Networks*. IEEE, May 2005, pp. 645–678.
- [2] —, *Clustering*. IEEE Press / Wiley, 2008.
- [3] A. Eiben and J. Smith, *Introduction to Evolutionary Computing*. Springer, 2007.
- [4] M. Illetskova, A. R. Bertels, J. M. Tuggle, A. Harter, S. Richter, D. R. Tauritz, S. Mulder, D. Bueno, M. Leger, and W. M. Siever, "Improving performance of cdcl sat solvers by automated design of variable selection heuristics," in *Proceedings of the 2017 IEEE Symposium Series on Computational Intelligence (SSCI 2017)*. IEEE, Nov. 2017.
- [5] A. S. Boragan and J. Fernandez-Villaverdei, "A comparison of programming languages in economics," in *Journal of Economic Dynamics and Control*. The National Bureau of Economic Research, Sep. 2014, pp. 265–273.
- [6] N. Hoda. Binaryheap: An implicit binary tree. [Online]. Available: <http://opendatastructures.org/versions/edition-0.1d/ods-java/node52.html>
- [7] C. D. Manning, P. Raghaven, and H. Schutze, *Introduction to Information Retrieval*. Cambridge University Press, 2008.
- [8] D. Dheeru and E. Karra Taniskidou, "Uci machine learning repository," 2017. [Online]. Available: <http://archive.ics.uci.edu/ml>
- [9] J. M. Tuggle. (2018) Kmeansgp. [Online]. Available: <https://github.com/tuglight/KMeansGP>