An Investigation into Comparing the Efficiency and Accuracy of Using The K-Means Clustering Algorithm at Sorting 50 Coins, From Ten Types, Using Their Weight and Diameter; Compared to a Human Sorting the Same Data Set

Primary Investigator: Jonathan Foot
Group Members: [OMMITED]
Tutor: [OMMITED]
The Scholar's Programme – UTC Reading – Year 12
5th Of April 2016

I. Abstract

The objectives of this investigation were to identify if the K-means clustering algorithm is more accurate and efficient at sorting a heterogeneous assortment of coins based on their weight and diameter properties; in comparison to a human sorting the same data set of coins physically by hand. To investigate the hypothesis an experiment was devised, which concluded the K-means clustering was more accurate and efficient than a human. The K-means clustering algorithm has an accuracy of 96% and a total elapsed time of 0.077612 seconds. Comparatively, the human test subjects had an average lower accuracy of 87.3% and a significantly slower time of 1:20.1706; these results supported the hypothesis.

II. Introduction

The problem statement is as follow: to design a K-means clustering machine learning algorithm which is capable of sorting different assortments of coins, with a greater accuracy and efficiency, compared to a human sorting the same dataset of coins, physically by hand.

Coin sorting has played a vital role in banks for many years. It is required to ensure the correct payment has been made and to prevent counterfeit coins in circulation. The method for which coins are sorted has evolved greatly over the years. Originally, coin sorting was accomplished by authorised professionals, who sorted coins manually by hand; this was slow and very labour intensive. This led to the development and utilisation of the first coin sorting devices in the late 1910s. These were immensely simplistic devices, composed of a tray containing a diverse range of different sized holes, which allowed the appropriate different sized denominations of coins to fall through into groupings. At the time this was considered, "ingenious" and sophisticated, boasting its ability to handle up to 60 coins at a time. [1]

Unfortunately, these devices were prone to having a large percentage error and slow sorting times, which led to the development of mechanising the process in the 1920s. Initial mechanised machines were primitive and statistically worse than the human counterparts; even in the 1960s money was still being counted by hand. It was only in the 1980s when technology advanced enough, that machine mechanisation took off. Now, coins could be counted at a rate of 72,000 per hour [2]. These sophisticated coin counting machines were first utilised on a large scale after the introduction of the Euro. Old currencies were substituted with the Euro, allowing charities the opportunity to collect in the pre-Euros to raise extra funds. The large volume and variety of coins preordained that physical sorting was not a practical option.

These machines would primarily focus on a coin's physical characteristics such as area, thickness and weight. This proved problematic when a homogenous collection of coins were inserted. The constraints of the machine limited their adaptability with sorting new and unseen before coins, resulting in a lot of inaccuracies [3].

An alternative approach is Machine Learning (a type of algorithm capable of learning from its mistakes, without the need to be explicitly programmed). Currently implemented examples would be search engines such as Google® and in services, such as, Netflix® which provide recommendations based on the user's previous choices and analytics [4]. Machine Learning can be further categorised into supervised and unsupervised learning; this investigation will be focusing on unsupervised. Unsupervised learning is input data containing no labelled responses; which is done to find hidden/ unknown groupings of data. [6]

The chosen machine learning algorithm used in this investigation is the K-means clustering algorithm. This will use the weight and diameter of coins to form them into distinct groupings called clusters [5]. The chosen characteristics weight and diameter were

chosen due to their correlation; generally, as the diameter of the coin increases so does its weight. Furthermore, the characteristics are easy to measure using a digital scale and a digital calliper respectively. A machine learning algorithms is a promising type of algorithms, capable of solving problems previously considered too difficult, or labour intensive, to solve by humans. Machine learning algorithms are already implemented with great success and this same method can be used to improve both the accuracy and efficiency of coin sorting.

The hypothesis for this investigation is as follows: The K-means clustering algorithm will have a greater accuracy and efficiency performing coin sorting using the coin's weight and diameter, in comparison to a human, sorting the same data set of coins.

The following hypothesis will be tested by comparing the time taken for a human to sort a stack of heterogeneous coins, in comparison to that of a k-means clustering machine learning algorithm. The sorting process will be graded on both the efficiency (the speed at which the task can be completed) and the accuracy (the percentage of successfully sorted coins).

III. Method

The purpose of the investigation is to distinguish if the K-means clustering algorithm is faster and more accurate at sorting a heterogeneous assortment of coins, in comparison to a human sorting the same data set. To determine this, a human will sort the coins physically by hand. The time taken to sort the coins into same denominations and the percentage error will be recorded and calculated. The process will then be repeated, but virtually, using a (machine learning) K-means clustering algorithm; which will use the coin's weight and diameter. A comparison and judgment will then be made between two sets of measurements: human and machine.

For the algorithm to function, properties of the coins must be collected including their weight and diameter. This will be recorded using a digital scale (an electronic device used to measure weight) and a digital calliper (a precision instrument, capable of measuring distances with extreme accuracy) respectively. In this investigation, a total of 50 coins of 10 types will be recorded. The recorded data characteristics can then be entered into the computer running Octave; the chosen programming language for this investigation due to being the de-facto standard in scientific research. It is these physical parameters of the coins that the computer will reference for sorting.

To conduct the algorithm, begin by selecting the start locations for the centroids; the start locations of the centroids can be arbitrary, but there should be the same number of centroids (a centroid is a data point at the centre of a cluster) as the number of types of coins. Next, calculate the Euclidian distance (the distances between two points of data) between every coin and each centroid using the equation:

$$Euclidian\ Distance = \sqrt{(x_i - x_{Centriod})^2 + (y_i - y_{centriod})^2}$$

$$x_i = X\ cordinate\ of\ the\ coin$$
$$x_{Centriod} = X\ cordinate\ of\ the\ centriod$$
$$y_i = y\ cordinate\ of\ the\ coin$$
$$y_{Centriod} = y\ cordinate\ of\ the\ centriod$$

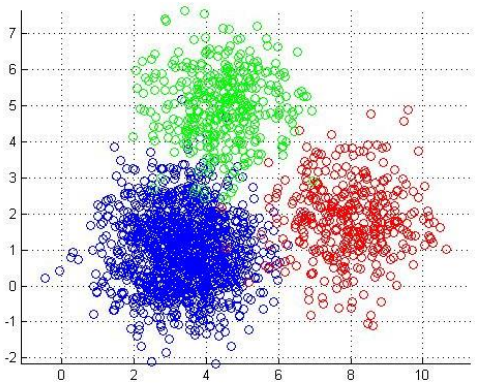
Using the Euclidian distances: distinguish which centroid each coin is nearest to. Now, each coin can be placed into a specific disjoint cluster grouping associated with that centroid; a cluster is a grouping of data that shares similar characteristics. Next, update the location of the centroids by calculating the mean weight and diameter of the coins in the cluster associated with that specific centroid. Using the equation:

$$Mean = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$N = The\ number\ of\ readings$$
$$\sum_{i=1}^{N} x_i = The\ Sum\ of\ all\ the\ readings$$

Using the new centroid location, repeat the algorithm to calculate new Euclidian distances and once again put the coins into clusters determined by which centroid they are nearest to. Iterate this process until the centroids remain stationary between the iterations; iteration is the process of repeating a segment of code, normally until some criteria has been meet. Once the algorithm has halted calculate the percentage error of the algorithm. Using the equation:

$$Percentage\ Error = \frac{Total\ Incorrectly\ placed\ coins}{Total\ number\ of\ coins} \times 100$$

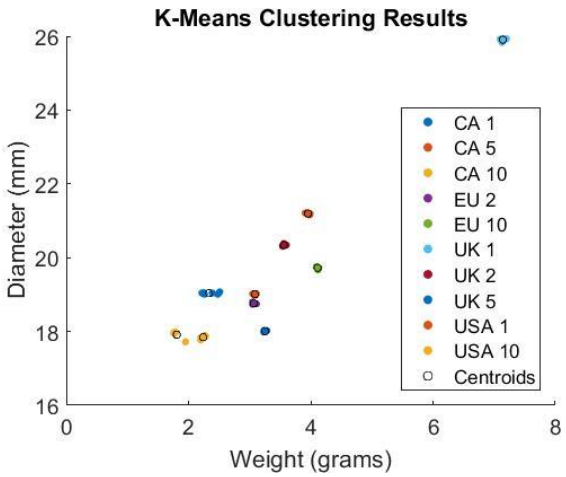


A Graphical Representation of K-means Clustering: Figure 1

Figure 1 shows a graphical representation of the K-means clustering algorithm. It shows 3 distinct clusters of data illustrated by the differences in colour. As previously stated each cluster is formed by finding the shortest Euclidian distance from a data point to a centroid. As such the figure shows there must have been 3 centroids to accompany the 3 clusters.

IV. Results

The results from the investigation are illustrated in Figure 2, 3 and 4.



Graph of the results: Figure 2

Mean Accuracy (Human):

$$\left(\left(\frac{40}{50}+\frac{50}{50}+\frac{41}{50}\right)\times 100\right) \div 3 = 87.3\%$$

Mean Completion Time (Human):

$$(57.524 + 92.743 + 90.245) \div 3 = 80.1706\text{s}$$

Accuracy (K-Means Clustering):

$$\frac{48}{50} \times 100 = 96\%$$

Time	Number correct	Total Coins	Accuracy
0:0.077612	48	50	96%

A table of results for K-means: Figure 3

Time	Number Correct	Total Coins	Accuracy
0:57.524	40	50	80%
1:32.743	50	50	100%
1:30.245	41	50	82%
Mean Time: 1:20.1706		**Mean Accuracy:** 87.3%	

A table of results for human trails: Figure 4

Figure 3 states the K-means algorithm had a completion time of only 0.077612 seconds and an accuracy of 96%. Comparatively, figure 4 states the average human time for completion was 1:20.1706 and an average accuracy of 87.3%. This shows that the K-means algorithm is both faster and more accurate at sorting the dataset of coins. Figure 2 is a graphical representation of the outcome of the K-means clustering algorithm, showing both the data points and the location of the centroids.

	Time	Incorrectly Placed coins
K-Means	25.87 Minutes	40,000
Human	18.56 Days	127,000

Hypothetical 1,000,000 coins - Figure 5

In this trial, there was a total of 50 coins of ten different types and from four different regions: Canada, Europe, UK and the USA. Using, these results a hypothetical prediction can be made for the outcome of sorting 1,000,000 coins. The K-means clustering algorithm would take approximately 1552.24 seconds or 25.87 Minutes and would make approximately 40,000 mistakes.

$$(1{,}000{,}000 \div 50) \times 0.077612 = 1552.24S$$
$$1552.24 \div 60 = 25.87Minutes$$

$$1{,}000{,}000 \times 0.004 = 40{,}000\ Coins$$

In reality, the true value for the time taken is likely to be greater, due to the algorithmic complexity of K-means clustering being an NP-hard solution [7]. NP-hard stands for "nondeterministic polynomial". Which connotates the worst-case running time to be a super-polynomial function. This meaning doubling the number of coins will not directly double the time elapsed due to them having a non-linear relationship but instead a polynomial relationship. Factors such as

number of centroids have a significant effect on the time and the program will be limited to specific computational/ hardware constraints of a device.

Comparatively, a human test subject would take approximately 1603412s or 18.5 days' continuous work and make approximately 127,000 mistakes.

$$(1{,}000{,}000 \div 50) \times 80.1706 = 1603412S$$
$$((160341 \div 60) \div 60) \div 24 = 18.56 \text{ Days}$$

$$1{,}000{,}000 \times 0.127 = 127{,}000\ Coins$$

Once again, this hypothetical prediction is likely to be inaccurate because a humans' accuracy should theoretically improve as they gain more experience sorting coins; learning from their mistakes and becoming more familiar with the physical characteristics and appearances of the different coins.

Interestingly, the only two incorrectly sorted coins in the K-means clustering algorithm were two USA one-cent coin, which were mistakenly identified for a Canadian one-cent coins. This was probably due to both coins being similar in weight and diameter.

V. Discussion/Conclusion

The hypothesis stated the K-means clustering algorithm will have a greater accuracy and efficiency sorting coins using their weight and diameter in comparison to a human sorting the same data set of coins. This hypothesis has been proven correct as illustrated in the results shown in figure 3 and figure 4. Which states the K-means clustering algorithm has an accuracy rating of 96%, whereas a human has a lower average of 87.3%. Additionally, the algorithm was significantly faster taking only 0.077612s, compared to the human average of 1:20.245. This definitively shows that the hypothesis is supported. Proving that machine learning is an ideal solution to coin sorting, due to its practicality, statistical advantages and ability to learn from mistakes.

While the hypothesis is proven successful, the accuracy and efficiency of the K-means algorithm could be further improved. Previous researchers have highlighted that the final clusters quality is heavily dependent upon the selection of the initial centroids and there is a high computational cost of finding the Euclidean distance on every iteration [8]. To improve the total elapsed time, you can reduce the number of calculations the algorithm makes on each iteration. This can be achieved by, finding the Euclidean distance from every data point and every centroid on the start iteration; placing them into their initial clusters. However, on each subsequent iteration only calculate the Euclidean distance between the data point and the centroid associated with its cluster. If this number is smaller than, or the same as before, then it must have remained inside that same cluster; eliminating the need to recalculate the Euclidian distance from every centroid. Reducing dramatically the number of calculations needed each iteration, especially if there is a large multitude of centroids. [10]

Another approach to reducing the algorithm's elapsed time is improving the preliminary centroid locations. If centroids are initially placed in a speculated optimal position the total amount of iterations needed for the centroids to remain stationary can be reduced. To generate an optimal start position, arrange the data in descending order and then split this data into groups dependent on the number of centroids. Next, find the average value of these groups to generate the start locations of centroids. [9] Reducing the number of iterations and the number of calculations needed in each iteration can significantly reduce the time taken especially when handling large datasets.

In addition to improving the efficiency, the accuracy could likewise be enhanced by considering additional physical properties of the coins, such as the number of sides it has. Measuring weight, diameter and number of sides on a coin allows a distinct character profile of a coin. For example, the new British £1 coin has 12 sides. This is a uniquely identifiable feature and could help distinguish it from coins with similar weights and diameters; preventing any incorrectly sorted coins. It is important to note, that adding a third dimension of measurement will increase the total elapsed time due to an increase in calculations needed. As such compromise must be made to evaluate if accuracy or efficiency is more imperative.

This investigation focused on unsupervised machine learning. A type of machine learning which uses data without any labelled responses. This method has been proven successful as it provides the ability to sort big data; data considered so large it was previously classified too large and complex to be analysed. Using K-means clustering machine learning, extremely large data sets can now be analysed computationally, revealing patterns and trends in data. This exemplifying the successfulness of machine learning.

In conclusion, the aim of the investigation was to discover if the K-means clustering algorithm was more accurate and efficient at sorting an heterogeneous assortment of coins compared to a human. The results proved the algorithm is both more efficient and accurate supporting the hypothesis. This is significant due to it highlighting the proficiencies and capabilities of using a machine learning algorithm. As previously stated, coin sorting has evolved greatly over the years and machine learning is a very credible forthcoming solution.

VI. Bibliography

[1] W.Kaempffert (ed.), "Counts and wraps coins quickly and accurately," Popular science monthly, vol. 94, no. 2, pp68, Feb. 1919.

[2] History of currency counting at the federal reserve bank of Philadelphia. (n.d.). federal reserve bank of Philadelphia. [Online].Available: https://www.philadelphiafed.org/education/teachers/resources/history-of-currency-counting Abbrev.March 14, 2017

[3] L.J.P. van der Maaten and P ,J. Boon, "Coin-O-Matic: A fast system for reliable coin classification," in Proceeding of the Muscle CIS Coin Competition, Vienna, 2006, pp. 7-18.

[4] Y.S. Abu-Mostafa, "Machines that think for themselves," Scientific American, vol. 307, no. 1, pp. 78-81, Jul. 2012

[5] Possibly the simplest way to explain K-Means algorithm. (27.May.2015). Bigdata-made simple,[online].Available: http://bigdata-madesimple.com/possibly-the-simplest-way-to-explain-k-means-algorithm. Accessed Abbrev March 4, 2017.

[6] What is machine learning? (7 July 2015). YouTube – Android Authority. [Online] Available. https://www.youtube.com/watch?v=WXHM_ifgGo&ab_channel=AndroidAuthority. Accessed. Abbrev March 2, 2017.

[7] "How Slow Is The K-Means Method?". (n.d.) Stanford University. [Online]. Available: http://theory.stanford.edu/~sergei/papers/kMeans-socg.pdf. Accessed Apr. 1, 2017.

[8] "Improving the Accuracy and Efficiency of the k-means Clustering Algorithm". (1 July 2009). Proceedings of the World Congress on Engineering. [Online]. Available: https://www.researchgate.net/profile/K_A_Nazeer/publication/44260003_Improving_the_Accuracy_and_Efficiency_of_the_k-means_Clustering_Algorithm/links/0fcfd51356e00827b8000000.pdf#. Accessed Apr. 1, 2017

[9] "Discovery of Preliminary Centroids Using Improved K- Means Clustering Algorithm". (2012). Department of Computer Science, Vignan university. [Online] Available: http://ijcsit.com/docs/Volume%203/vol3Issue3/ijcsit20120303151.pdf .Accessed Apr.1, 2017

[10] "An efficient enhanced k-means clustering algorithm". (11 May 2006). Suez canal university, Ain Shams university, Minutia university. [Online]. Available: https://www.researchgate.net/publication/226683377_Efficient_enhanced_k-means_clustering_algorithm .Accessed Apr.1, 2017