Improving the Cluster Efficiency on Sea Level Rise Dataset using Data Discretization

Sharon Dominick
Assistant Professor
Bishop Heber College (Autonomous)
Tiruchirappalli - 17

T. Abdul Razak, Ph.D
Associate Professor
Jamal Mohamed College (Autonomous)
Tiruchirappalli - 20

ABSTRACT
Rising sea levels, an effect of global warming, is a cause of concern and it is likely to affect the developing countries. With respect to the data set published for research at the World Bank, clustering a data mining technique is applied to detect the most likely to be affected regions. When tested with the k-Means clustering technique, the result of the clustering process reveals a lot of imperfections; this research analyzes the use of data discretization to improve the quality of the clustering process.

Keywords
Discretization, Clustering, Partitioning, vulnerable.

1. INTRODUCTION
Global warming, which caused the rise in average temperature of Earth's atmosphere and oceans is projected to continue. The effects include rise in sea level, a change in the amount and pattern of precipitation, a probable expansion of subtropical deserts. Other effects of the warming include a more frequent occurrence of extreme-weather events including heat waves, droughts and heavy rainfall, ocean acidification and species extinctions due to shifting temperature regimes. Effects significant to humans include the threat to food security from decreasing crop yields and the loss of habitat from inundation. Climate change is likely to adversely affect millions of people through increased coastal flooding [1], reductions in water supplies, increased malnutrition and increased health impacts. Most economic studies suggest losses of world gross domestic product (GDP) for this.

Melting glaciers and land-based ice sheets also contribute to rising sea levels, threatening low-lying areas around the globe with beach erosion, coastal flooding, and contamination of freshwater supplies [2]. Small islands as a result of sea level rise are expected to threaten vital infrastructure and human settlements, causing homelessness in countries with low lying areas.

This research aims at identifying the vulnerable zones, for which the k- Mean clustering technique of data mining is used.

1.1 Motivation
This research draws its inspiration from a recent global problem, global warming. Its effect of rising sea levels is discussed here; this research is based on the collective data summarized by Susmita Dasgupta [3].

1.2 Research Objective
The objective of this research is to study the impact of rise in sea level, on developing countries using data mining techniques.

2. RELATED WORK
Alsabti et al. [4] proposed the use of a k-d tree structure to find the patterns close to the given prototype.

Fang-Xiang et al. [5] in a new method suggested the use of both Genetic Algorithms and Iterative Optimal K-Means Algorithm (IOKMA), which used only three genetic operators, selection, mutation and crossover.

Hautam¨aki et al. [6] suggested a two step process by which the outliers were removed and then clustered.

Arthur et al. [7] suggested the use of randomized seeding to improve the speed and accuracy of the K-Means clustering technique.

K.Arun Prabha et al. [8] proposed the use of Genetic Algorithms to increase the performance of clustering.

3. CLUSTERING
Clustering is the process of grouping a set of data objects that are similar to one another within the same cluster and are dissimilar to objects in other clusters.

Clustering is an example of unsupervised learning, it is a form of learning by observation rather than learning by examples. The advantages of clustering based process is that it is adaptable to changes and helps single out useful features that distinguish different groups. The k-means clustering technique is applied to the chosen data set [9].

3.1 Clustering by K-Means Partitioning
Algorithm: k-means. Here each cluster's center is represented by the mean value of the objects in the cluster. [9]

Input:

k: the number of clusters

D: a data set containing n objects.

Output: A set of clusters**:**

i. Arbitrarily choose k objects from D as the initial cluster centers;

ii. Repeat

a. (re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;

b. update the cluster means i.e., calculate the mean value of the objects for each cluster;

iii. until no change;

3.2 Observations of K-means clustering

The results of the clustering process reveal a huge number of incorrectly clustered instances. In order to increase the efficiency of the clustering process, a technique known as data discretization is applied to the dataset.

3.2.1 Data Discretization

Discretization refers to the process of converting or partitioning continuous attributes, features or variables to discretized or nominal attributes or features or variables or intervals. Typically data is discretized into partitions of K equal lengths/width (equal intervals) or K% of the total data (equal frequencies).[10]

4. IMPLEMENTATION

The research uses WEKA and is based on the dataset by Susmita Dasgupta published for research at the World Bank website [3]. The dataset to be used in WEKA is in an .arff or .csv file. The dataset, "SeaLevelRise.arff" consists of 504 instances with 15 attributes. The attributes denote the Country Code, Country Name, Zone, Country Area(in SqKm), Scenario (which may be Land, Population, GDP, Agriculture, Urban Extent and Wetland), 1 meter, 2 meter, 3 meter,4 meter,5 meter, 1 meter impact, 2 meter impact, 3 meter impact, 4 meter impact, 5 meter impact. The dataset is preprocessed, K-Means clustering technique is applied, and its effect is studied.

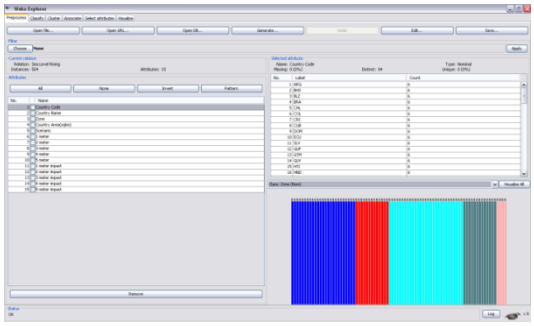


Fig 1: The imported dataset with its attributes.

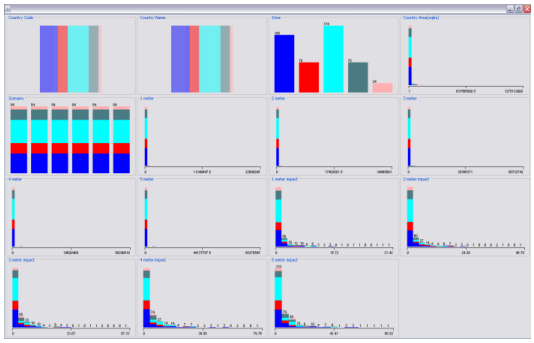


Fig 2: Visualization of the entire data set with respect to its attributes.

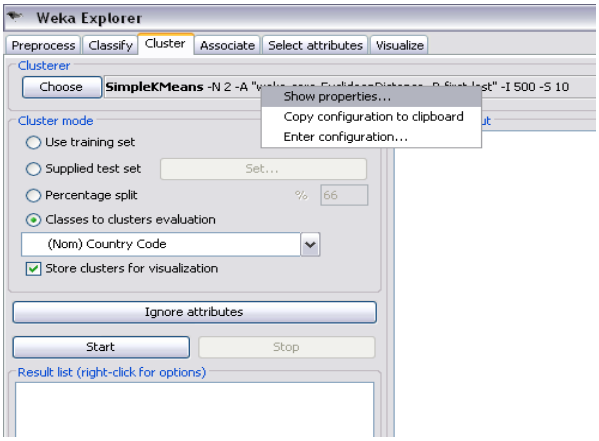


Fig 3: The cluster option is selected and the properties of the clustering technique are altered.

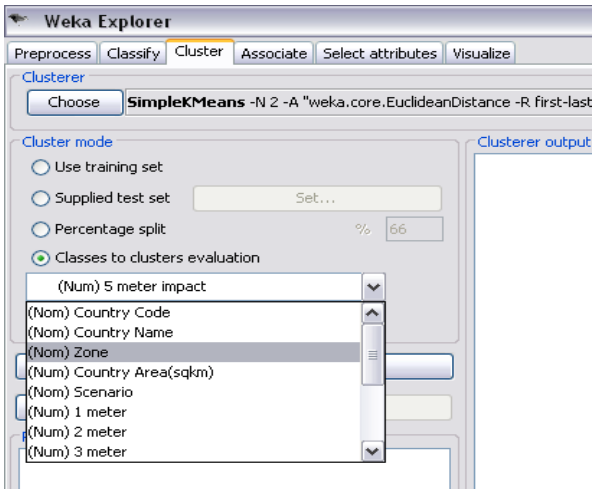


Fig 4: The dataset is chosen to be clustered based on the "Zone" attribute.

The output is generated and the dataset is clustered. The above clustering process has resulted in clustering the dataset into 5 clusters, but with 72% of the data being incorrectly clustered.

In order to reduce the % of incorrectly clustered data, we use "Data Discretization" and filter technique. The unsupervised learning method is chosen to discretize the attributes.

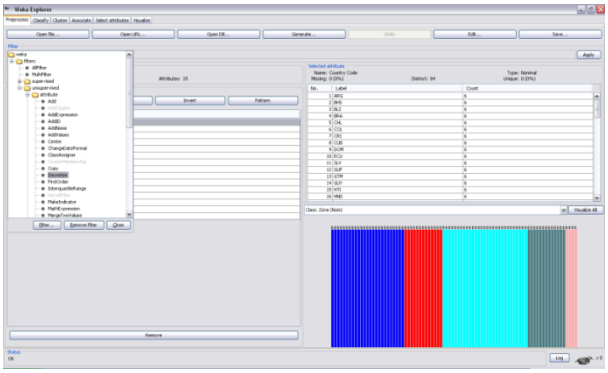


Fig 5: The "Discretize" option is chosen.

Since the discretization method has been applied to the data set the appearance of the data set appears different from its original view before discretization.

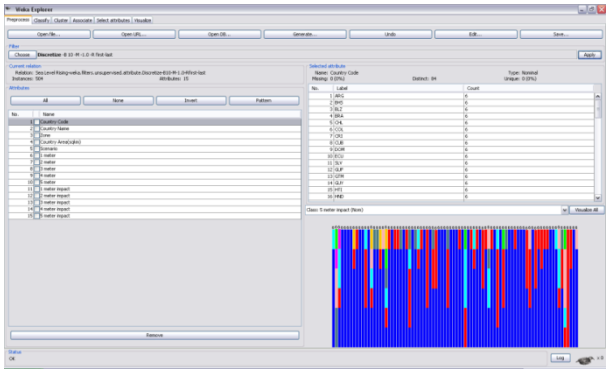


Fig 6: The representation of the dataset, after applying discretization.

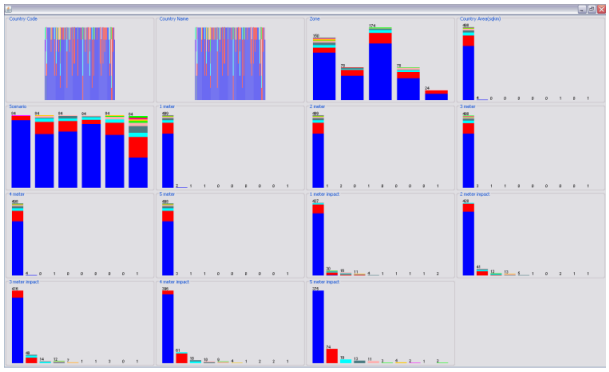


Fig 7: The representation of the various attributes in the dataset post Discretization.

The dataset is clustered using the same K-Means Clustering Technique. But this time, the % of incorrectly clustered data is significantly less. It's about 66%.

```
Clustered Instances

0       52 ( 10%)
1       21 (  4%)
2       12 (  2%)
3      195 ( 39%)
4      224 ( 44%)


Class attribute: Zone
Classes to Clusters:

  0  1  2  3  4  <-- assigned to cluster
 16  7  5 56 66 | Latin America / Caribbean
 13  4  1 28 32 | Middle East / North Africa
 18  4  3 56 93 | Sub-Saharan Africa
  5  5  3 42 23 | East Asia
  0  1  0 13 10 | South Asia

Cluster 0 <-- Middle East / North Africa
Cluster 1 <-- East Asia
Cluster 2 <-- No class
Cluster 3 <-- Latin America / Caribbean
Cluster 4 <-- Sub-Saharan Africa

Incorrectly clustered instances :       337.0     66.8651 %
```

Fig 8: The % of incorrectly clustered after applying discretization.

It is noted that the efficiency of the clustering is less, when compared to the clustering done after data discretization.

5. EXPECTED RESULTS

Table 1: Clustering done without data being discretized.

Clustering before applying Discretization		
Cluster	**No of Clustered Instances**	**Percentage of Clustering**
0	114	23
1	107	21
2	28	6
3	170	34
4	85	17

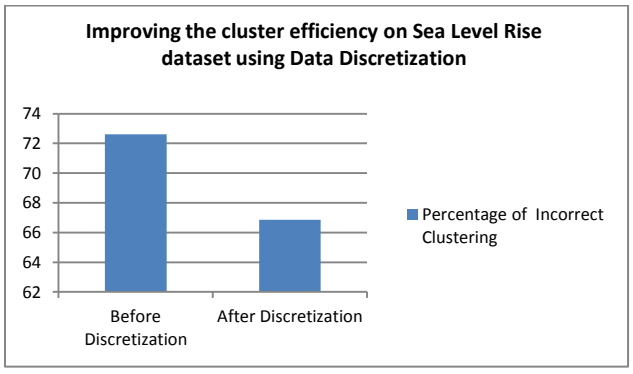
Table 2: Clustering done after data being discretized.

Clustering after applying Discretization		
Cluster	**No of Clustered Instances**	**Percentage of Clustering**
0	52	10
1	21	4
2	12	2
3	195	39
4	224	44

Table 3: Improvement in the clustering process.

Status	No of Incorrectly Clustered Instances	Percentage of Clustering
Before Discretization	366	72.619
After Discretization	337	66.8651

Chart 1: Decrease in in-correct clustering percentage.



There is significant improvement in the clustering process, as the numbers of incorrectly clustered instances have been reduced. Data Discretization has helped achieve this.

6. CONCLUDING REMARKS

It is widely known that machine learning algorithms produce better models by discretization of continuous attributes. This work reflects on the usage of discretization, a technique of placing the values in buckets, to limit the number of possible states. The buckets are treated as ordered and discrete values. The K-Means Clustering, a machine learning algorithm was

applied to a test dataset and the results showed significant improvement while using discretizaton techniques.

Based on the results of the discretizaton process, which reduces the incorrectly clustered instances and improves clustering, any predictions made using the data will prove to be more precise.

This would help; identify the most vulnerable zones, prone to sea level rise, better.

7. ACKNOWLEDGMENTS

As the pole star guides every lonely traveller with assurance and safety, I extend my sincere gratitude to my guide Dr. T. Abdul Razak, for his guidance, encouragement, motivation and inspiration. I convey my heartfelt thanks to my beloved family and friends, who helped in the successful completion of the project. Finally I am very thankful to all who supported me, in the completion of my research.

8. REFERENCES

[1] http://www.nrdc.org/globalwarming/qthinice.asp.

[2] http://www.ucsusa.org/global_warming/science_and_impacts/impacts/infographic-sea-level-rise-global-warming.html.

[3] http://econ.worldbank.org/WBSITE/EXTERNAL/EXTDEC/EXTRESEARCH/0,,contentMDK:22185588~pagePK:64214825~piPK:64214943~theSitePK:469382,00.htm.

[4] Alsabti, Khaled; Ranka, Sanjay; and Singh, Vineet, "An efficient k-means clustering algorithm". Electrical Engineering and Computer Science.1993.

[5] Fang-Xiang Wu, W.J.Zhang and Anthony J. Kusalik,"A Genetic K-Means Clustering Algorithm applied to Gene Expression Data", Springer-Verlag Berlin Heidelberg.2003.

[6] Ville Hautam¨aki, Svetlana Cherednichenko, Ismo K¨arkk¨ainen,Tomi Kinnunen, and Pasi Fr¨anti,"Improving K-Means by Outlier Removal", Springer-Verlag Berlin Heidelberg. 2005.

[7] Arthur, D. and Vassilvitskii, S. "k-means++: the advantages of careful seeding". Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, Philadelphia, USA. 2007.

[8] Arun Prabha,R.Saranya,"Refinement of K-Means Clustering using Genetic Algorithm", Journal of Computer Applications(JCA),2011.

[9] Jiawei Han, Micheline Kamber, and Jian Pei,"Data mining Concepts and Techniques", 2nd ed, Morgan Kaufmann,2006.

[10] http://www.improvedoutcomes.com/docs/WebSiteDocs/Classification_and_Prediction/SLAM/Discretization.htm

[11] http://www.nrdc.org/globalwarming/fcons/fcons4.asp.

9. AUTHOR PROFILE

Mrs. Sharon Dominick is working as an Assistant Professor in Department of Computer Applications, Bishop Heber College (Autonomous), Tiruchirappalli, Tamil Nadu, India. She has 1 year experience in teaching field. Her current area of research is Data Mining.

Dr. T. Abdul Razak is working as an Associate Professor in Department of Computer Science, Jamal Mohamed College (Autonomous), Tiruchirappalli, Tamil Nadu, India. He has more than 26 years of experience in teaching field. His areas of interest include Microprocessors, Digital Electronics, Computer Organization and Architecture and Data Mining. His current area of research is Data Mining.