RESEARCH ARTICLE OPEN ACCESS

Mineral Detection using K-Means Clustering Technique

P. Bangarraju[1], R. V. S. Manohar[2], Y. L. Anusha[3], Y. Suresh[4], B. Ram Mohan[5]
(Asst. Professor)
Lendi Institute of Engineering and Technology, Electronics And Communication Engineering, Jonnada, Vizianagaram, INDIA, Pin-535002.

Abstract
This paper is all about a novel algorithm formulated with k-means clustering performed on remote sensing images. The fields of Remote Sensing are very wide and its techniques and applications are used both in the data acquisition method and data processing procedures. It is also a fast developing field with respect to all the above terms. Remote Sensing plays a very important role in understanding the natural and human processes affecting the earth's minerals. The k-means clustering technique is used for segmentation or feature selection of passive and active imaging and non-imaging Remote Sensing, on airborne or on satellite platforms, from monochromatic to hyperspectral. So here we concentrate on the images taken on or above the surface of the earth which are applied based on the proposed algorithm to detect the minerals like Giacomo that exist on the surface of the earth. Our experimental results demonstrate that our technique can improve the computational speed of the direct k-means algorithm by an order to two orders of magnitude in the total number of distance calculations and the overall time.

Key words: K-means clustering, sobel edge detection, remote sensing, hyperspectral image, feature extraction.

I. INTRODUCTION

Clustering is the process of partitioning or grouping a given set of patterns into disjoint *clusters*. This is done such that patterns in the same cluster are alike and patterns belonging to two different clusters are different. Clustering has been a widely studied problem in a variety of application domains including neural networks. Several algorithms have been proposed in the literature for clustering.

The k-means method has been shown to be effective in producing good clustering results for many practical applications. However, a direct algorithm of k-means method requires time proportional to the product of number of patterns and number of clusters per iteration. This is computationally very expensive especially for large datasets. We propose a novel algorithm for implementing the kmeans method. Our algorithm produces the same or comparable (due to the round-off errors) clustering results to the direct k-means algorithm. It has significantly superior performance than the direct k-means algorithm in most cases. The rest of this paper is organized as follows. We review previously proposed approaches for improving the performance of the k-means algorithms in Section 2. We present our algorithm in Section 3. We describe the experimental results in Section 4 and we conclude with Section 5.

II. REMOTE SENSING

1. The Definition of **Remote Sensing** In the broadest sense, the measurement or acquisition of information of some property of an object or phenomenon, by a recording device that is not in physical or intimate contact with the object or phenomenon under study; e.g., the utilization at a distance (as from aircraft, spacecraft, or ship) of any device and its attendant display for gathering information pertinent to the environment, such as measurements of force fields, electromagnetic radiation, or acoustic energy. The technique employs such devices as the camera, lasers, and radio frequency receivers, radar systems, sonar, seismographs, gravimeters, magnetometers, and scintillation counters. The template is designed so that author affiliations are not repeated each time for multiple authors of the same affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization). This template was designed for two affiliations.
2. The practice of data collection in the wavelengths from ultraviolet to radio regions. This restricted sense is the practical outgrowth from airborne photography. Sense (1) is preferred and thus includes regions of the EM spectrum as well as techniques traditionally considered as belonging to conventional geophysics.

- Today, we define satellite remote sensing as the use of satellite-borne sensors to observe, measure, and record the electromagnetic radiation reflected or emitted by the Earth and its environment for subsequent analysis and extraction of information.

II.A. REMOTE SENSING *VS* AERIAL PHOTOGRAPHY / PHOTOGRAMMETRY:

Both systems gather data about the upper surface of the Earth, by measuring the Electromagnetic radiation, from airborne systems. The following major differences can be given:

- Aerial photos are taken by an analog instrument: a film of a (photogrammetric) camera, then scanned to be transformed to digital media. Remote Sensing data is usually gathered by a digital CCD camera.
- An Aerial photograph is a central projection, with the whole picture taken at one instance. A Remote Sensing image is created line after line; therefore, the geometrical correction is much more complex, with each line (or even pixel) needing to be treated as a central projection.
- Aerial photos usually gather data only in the visible spectrum (there are also special films sensitive to near infrared radiation), while Remote Sensing sensors can be designed to measure radiation all along the Electromagnetic spectrum.
- Aerial photos are usually taken from planes, Remote Sensing images also from satellites.
- Both systems are affected by atmospheric disturbances. Aerial photos mainly from haze (that is, the scattering of light – the process which makes the sky blue), Remote Sensing images also from processes of absorption. Atmospheric corrections to Aerial photos can be made while taking the picture (using a filter), or in post-processing, as in done Remote Sensing. Thermal Remote Sensing sensors can operate also at nighttime and Radar data is almost weather independent.
- In Photogrammetry the main efforts are dedicated for the accurate creation of a 3d model, in order to plot with high accuracy the location and boundaries of objects, and to create a Digital Elevation Model, by applying sophisticated geometric corrections. In Remote Sensing the main efforts are dedicated for the analysis of the incoming Electromagnetic spectrum, using atmospheric corrections, sophisticated statistical methods for classification of the pixels to different categories, and analysing the data according to known physical processes that affect the light as it moves in space and interacts with objects.
- Remote Sensing images are very useful for tracking phenomena on regional, continental and even global scale, using the fact that satellites cover in each image a wide area, and taking images all the time (whether fixed above a certain point, or “revisiting” the same place every 15 days (for example).
- Remote Sensing images are available since the early 1970’s. Aerial photos, provide a longer time span for landscape change detection (the regular coverage of Israel by Aerial photos started in 1944/5, for example, with many Aerial photos taken also during World War 1).
- Remote Sensing images are more difficult to process, and require trained personnel, while aerial photographs can be interpreted more easily.

II.B. REMOTE SENSING VS SONAR

The SONAR can also be considered as Remote Sensing – that is, studying the surfaces of the sea (bathymetry and sea bed features) from a distance. The SONAR is an active type of Remote Sensing (like Radar; Not depending on an external source of waves, measuring the time between the transmission and reception of waves produced by our instruments, and their intensity), but using sound waves, and not Electromagnetic radiation.

Both systems transmit waves through an interfering medium (water, air), that adds noise to the data we are looking for, and there for corrections must be applied to the raw data collected. In Remote Sensing however, Radar is considered to be almost weather independent, and atmospheric disturbances affect mainly passive Remote Sensing). To make these necessary corrections, both systems depend on calibration from field data (be it salinity, temperature and pressure measured by the ship while surveying, or measurements of the atmospheric profile parameters by a meteorological radiosonde for example). Sonar’s are mainly used to produce the bathymetry of the sea, while Remote Sensing techniques are focusing more on identification of the material’s properties than on its height.

III. ELECTROMAGNETIC ENERGY:

Electromagnetic energy refers to all energy that moves with the velocity of light in a harmonic wave pattern. The word *harmonic* implies that the component waves are equally and repetitively spaced in time. The wave concept explains the propagation of Electromagnetic energy, but this energy is detectable only in terms of its interaction with matter. In this interaction, Electromagnetic energy behaves as though it consists of many individual bodies called *photons* that have such particle-like properties as energy and momentum.

Electromagnetic waves can be described in terms of their:

· Velocity: The speed of light, c=3*108 m*sec^{-1}).

· Wavelength: l, the distance from any position in a cycle to the same position in the next cycle, measured in the standard metric system. Two units are usually used: the micrometer (mm, 10-6m) and the nanometer (nm, 10-9m).

· Frequency: n, the number of wave crests passing a given point in specific unit of time, with one hertz being the unit for a frequency of one cycle per second. Wavelength and frequency are related by the following formula:

$$\boxed{c = 1 * n}$$

Electro-Magnetic radiation consists of an electrical field (E) which varies in magnitude in a direction perpendicular to the direction in which the radiation is traveling, and a magnetic field (M) oriented at right angles to the electrical field. Both these fields travel at the speed of light (c).

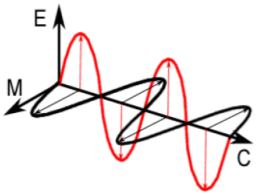


Figure: Electro-Magnetic radiation

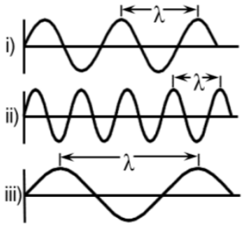


Figure: wavelength and frequency

III.A. MECHANISM:

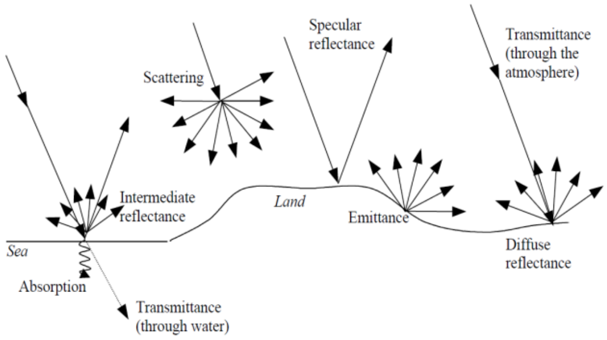
□ Radiation may be *transmitted*, that is, passed through the substance. The velocity of Electromagnetic radiation changes as it is transmitted from air, or a vacuum, into other substances.

□ Radiation may be*absorbed* by a substance and give up its energy largely to heating the substance.

□ Radiation may be *emitted* by a substance as a function of its structure and temperature. All matter at temperatures above absolute zero, 0°K, emits energy.

□ Radiation may be*scattered*, that is, deflected in all directions and lost ultimately to absorption or further scattering (as light is scattered in the atmosphere).

□ Radiation may be *reflected*. If it is returned unchanged from the surface of a substance with the angle equal and opposite to the angle of incidence, it is termed *specular* reflectance (as in a mirror). If radiation is reflected equally in all directions, it is termed *diffuse*. Real materials lie somewhere in between.



The interactions with any particular form of matter are selective with regard to the Electromagnetic radiation and are specific for that form of matter, depending primarily upon its surface properties and its atomic and molecular structure.

IV. PLANCK'S RADIATION LAW

The primary law governing blackbody radiation is the *Planck Radiation Law*, which governs the intensity of radiation emitted by unit surface area into a fixed direction (solid angle) from the blackbody as a function of wavelength for a fixed temperature. The Planck Law can be expressed through the following equation.

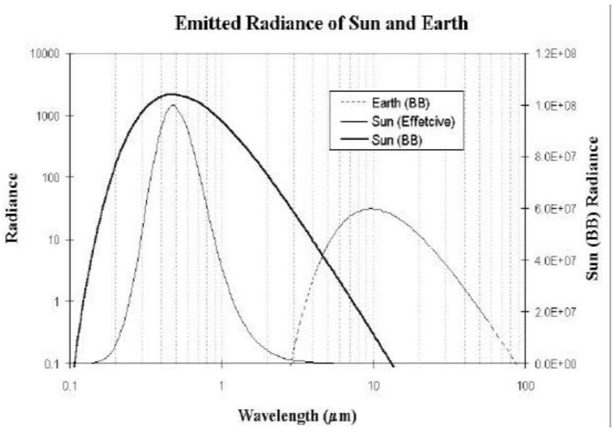
$$E(\lambda, T) = \frac{2hc^2}{\lambda^5} \frac{1}{e^{hc/\lambda kT} - 1}$$

$$h = 6.625 \times 10^{-27} \text{ erg-sec} \quad \text{(Planck Constant)}$$

$$k = 1.38 \times 10^{-16} \text{ erg/K} \quad \text{(Boltzmann Constant)}$$

$$c = 3 \times 10^{10} \text{ cm/sec} \quad \text{(Speed of Light)}$$

Every object with a temperature above the absolute zero radiates energy. The relationship between wavelength and the amount of energy radiated at different wavelengths, is shown in the following figure, and formulated above.

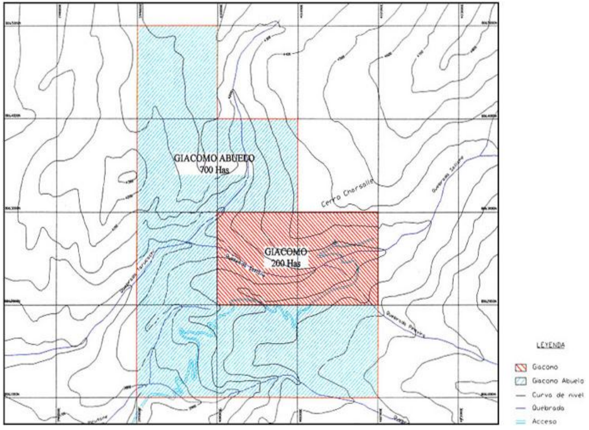


V. GIACOMO MINERAL DEPOSIT:

In the Southern Peruvian Andes, a unique mineral occurrence of Silicon Dioxide (SiO2) and Titanium Dioxide (TiO2) exists near the city of Tacna. This occurrence is known as the Giacomo Project. Giacomo is currently estimated to contain a resource of at least 100 million metric tonnes of high-grade SiO2 and TiO2 mineralization of extremely small particle size.

Giacomo is really more aptly described as an "Occurrence" rather than a Deposit. The white outcroppings seen below are actually the minerals of interest. The following pictures reveal the purity of the underlying material.

The mine was discovered 10 years ago and is owned by a single individual. The project started at the time of its discovery, however because of the owner's ill health it was paused until now. The project is looking for commercial development agreements, to assign the rights for the exploitation, commercialization and direction of the project. Currently there is no agreement with any utility company or a specific customer. The deposit has never been exploited and has been preserved in the state originally found, except for analysis-oriented excavations. However, due to the deposit's characteristics as well as made investments; it is currently ready to generate cash flow.



The Giacomo Deposit is located in the South of Peru, in the district of Estique in the province of Tarata, in the Tacna region. Estique is approximately 60 kilometres northeast of the City of Tacna. The concession is on the western foothills of the Barroso's Mountain Chain at an average altitude of 4,500 MASL.

VI. SOBEL EDGE DETECTION

Sobel which is a popular edge detection method is considered in this work. There exists a function, edge.m which is in the image toolbox. In the edge function, the Sobel method uses the derivative approximation to find edges. Therefore, it returns edges at those points where the gradient of the considered image is maximum. The horizontal and vertical gradient matrices whose dimensions are 3×3 for the Sobel method has been generally used in the edge detection operations. In this work, a function is developed to find edges using the matrices whose dimensions are 5×5 in matlab. Standard Sobel operators, for a 3×3 neighborhood, each simple central gradient estimate is vector sum of a pair of orthogonal vectors [1]. Each orthogonal vector is a directional derivative estimate multiplied by a unit vector specifying the derivative's direction. The vector sum of these simple gradient estimates amounts to a vector sum of the 8 directional derivative vectors. Thus for a point on Cartesian grid and its eight neighbors having density values as shown:

a	b	c
d	e	f
g	h	i

In [1], the directional derivative estimate vector *G* was defined such as density difference / distance to neighbor. This vector is determined such that the direction of *G* will be given by the unit vector to the approximate neighbor. Note that, the neighbors group into antipodal pairs: (a,i), (b,h), (c,g), (f,d). The vector sum for this gradient estimate: Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

$$G = \frac{(c-g)}{R} \cdot \frac{[1,1]}{R} + \frac{(a-i)}{R} \cdot \frac{[-1,1]}{R} + (b-h) \cdot [0,1] + (f-d) \cdot [1,0]$$

Where, $R = 2$. This vector is obtained as:

$$G = [(c-g-a+i)/2 + f - d,\ (c-g+a-i)/2 + b - h]$$

Here, this vector is multiplied by 2 because of replacing the divide by 2. The resultant formula is given as follows (see, for detail [1]):

$$G' = 2.G = [(c - g - a + i) + 2.(f - d),\ (c - g + a - i) + 2.(b - h)]$$

The following weighting functions for x and y components were obtained by using the above vector.

1	0	1
-2	0	2
-1	0	1

1	2	1
0	0	0
-1	-2	-1

Now, we explain that the dimensions of the matrices are extended by using [1]. The definition of the gradient can be used for 5×5 neighborhood [8]. In this case, twelve directional gradients must be determined instead of four gradients. The following figure 5×5 neighborhood.

a	b	c	d	e
f	g	h	i	j
k	l	m	n	o
p	r	s	t	u
v	w	x	y	z

The resultant vector *G'* (similar to the determination of Sobel 3×3 method) for 5×5 is given as follows:

$$\begin{aligned} G' = [&20(n-l) + 10(i - r - g + t + o - k) + 5(e - v - a + z) + 4(d - w - b + y) \\ &+ 8(j - p - f + u), 20(h - s) + 10(i - r + g - t) + 5 \cdot (e - v + a - z) \\ &+ 4(j - p + f - u) + 8(d - w + b - y)] \end{aligned}$$

The horizontal and vertical masks are obtained by using the coefficients in this equation such as [8]

-5	-4	0	4	5
-8	-10	0	10	8
-10	-20	0	20	10
-8	-10	0	10	8
-5	-4	0	4	5

5	8	10	8	5
4	10	20	10	4
0	0	0	0	0
-4	-10	-20	-10	-4
-5	-8	-10	-8	-5

These masks are used by the edge detection function in the following section.

VII. K-MEANS CLUSTERING

A. *K-means clustering effectiveness.*

The k-means method has been shown to be effective in producing good clustering results for many practical applications. However, a direct algorithm of k-means method requires time proportional to the product of number of patterns and number of clusters per iteration. This is computationally very expensive especially for large datasets. We propose a novel algorithm for implementing the kmeans method. Our algorithm produces the same or comparable (due to the round-off errors) clustering results to the direct k-means algorithm. It has significantly superior performance than the direct k-means algorithm in most cases. The rest of this paper is organized as follows. We review previously proposed approaches for improving the performance of the k-means algorithms in Section 2. We present our algorithm in Section 3. We describe the experimental results in Section 4 and we conclude with Section 5.

The template is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin in this template measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

VIII. PROPOSED ALGORITHM

In this section, we briefly describe the direct k-means algorithm [9, 8, 3]. The number of clusters *k* is assumed to be fixed in k-means clustering. Let the *k* prototypes $(w_1, \ldots, w_k)$ be initialized to one of the *n* input patterns $(i_1, \ldots, i_n)$. Therefore,

$$w_j = i_l,\ j \in \{1, \ldots, k\},\ l \in \{1, \ldots, n\}$$

Figure 1 shows a high level description of the direct kmeans clustering algorithm. / _ is the _th cluster whose value is a disjoint subset of input patterns. The quality of the clustering is determined by the following error function:

$$E = \sum_{j=1}^{k} \sum_{i_l \in C_j} \mid i_l - w_j \mid^2$$

The appropriate choice of *k* is problem and domain dependent and generally a user tries several values of _ Assuming that there are *n* patterns, each of dimension @, the computational cost of a direct k-means algorithm per iteration (of the repeat loop) can be decomposed into three parts:

1. The time required for the first *for* loop in Figure 1 is $O(nkd)$.

2. The time required for calculating the centroids (second *for* loop in Figure 1) is $O(nd)$.

3. The time required for calculating the error function is

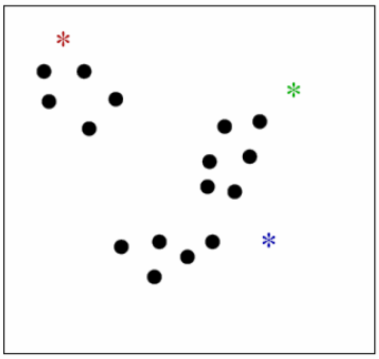
$O(nd)$.

The number of iterations required can vary in a wide range from a few to several thousand depending on the number of patterns, number of clusters, and the input data distribution. Thus, a direct implementation of the k-means method can be computationally very intensive. This is especially true for typical data mining applications with large number of pattern vectors.

B. Our Algorithm

- K = # of clusters (given); one “mean” per cluster
- Interval data
- Initialize means (e.g. by picking k samples at random)
- Iterate:

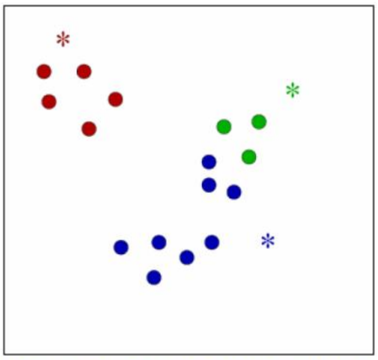
(1) assign each point to nearest mean
(2) move “mean” to center of its cluster.



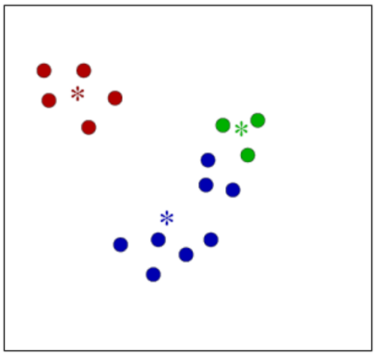
Initialize representatives (“means”)

□. □. □. □̃1 .1

(1) Assignment Step; Means



Assign to nearest representative

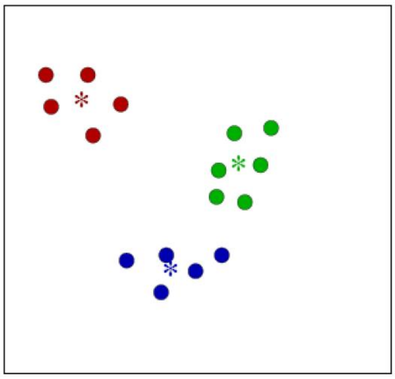


Re-estimate means

Convergence after another iteration
Complexity:
O(k . n . # of iterations

The Objective Function is:

$$\min_{\{\boldsymbol{\mu}_1,\cdots,\boldsymbol{\mu}_k\}} \sum_{h=1} \sum_{\mathbf{x}\in\mathcal{X}_h} \|\mathbf{x}-\boldsymbol{\mu}_h\|^2$$



IX. K-MEANS CLUSTERING – DETAIL

- Complexity is O(n * K * I * d)
  - ➢ n = number of points, K = number of clusters,
  - ➢ I = number of iterations, d = number of attributes
  - ➢ Easily parallelize
  - ➢ Use kd-trees or other efficient spatial data structures for some situations
    - ❖ Pelleg and Moore (X-means)
- Sensitivity to initial conditions
- A good clustering with smaller K can have a lower SSE than a poor clustering with higher K

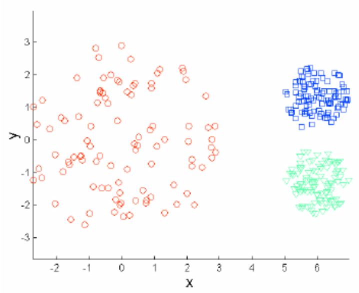
X. LIMITATIONS OF K-MEANS CLUSTERING

- K-means has problems when clusters are of differing
  – Sizes

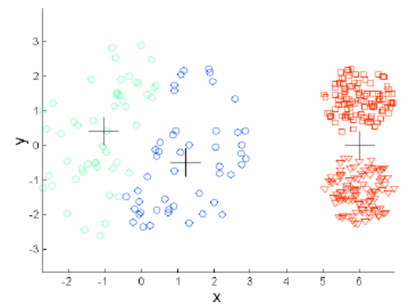
– Densities
– Non-globular shapes

- Problems with outliers
- Empty clusters

XI. Limitations of K-means: Differing Density

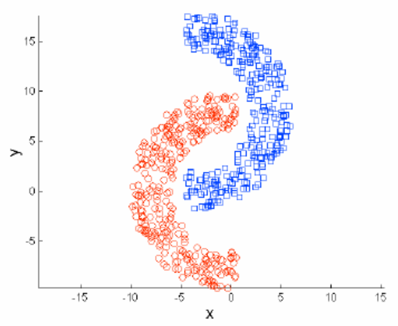


Original Points

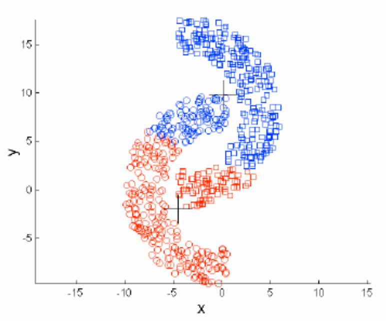


K-means (3 Clusters)

XII. Limitations of K-means: Non-globular Shapes

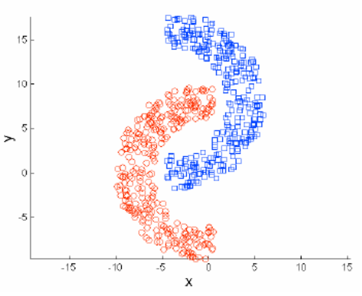


Original Points

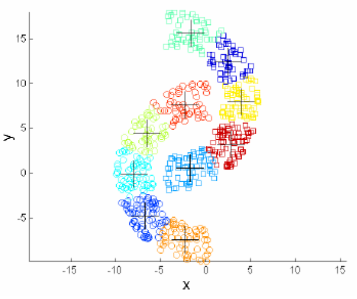


K-means (2 Clusters)

XIII. Overcoming K-means Limitations



Original Points



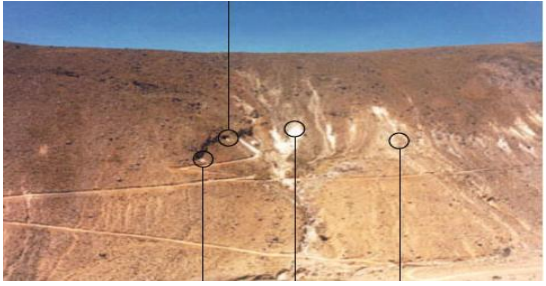
K-means Clusters

XIV. RESULT

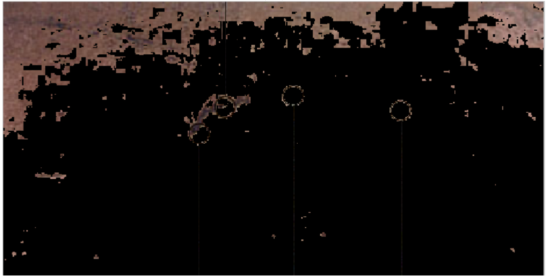
The proposed algorithm has been applied to images containing area where giacomo mineral is existed under earth's surface.

As shown in the fig(i), the image is taken from a hill area containing giacomo mineral. Giacomo is generally whitish mineral and here we are considering spectral colors emitting from the image, so here the maximum percentage of color emitted is blue, then we apply k-means clustering followed by edge detection, which highlights the area containing giacomo mineral, the border of that area. These two images added together which gives the mineral extraction and border identification. The resultant figure shown in the fig(ii). Similarly, the fig(iii) as feature extracted shows the areas of the giacomo mineral is given in the fig(iv).

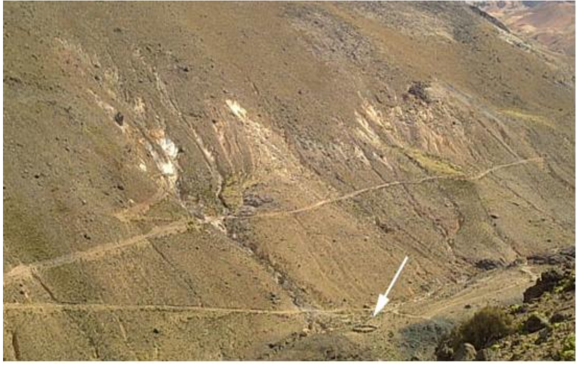
(i)



(ii)



(iii)



(iv)

XV. CONCLUSION

The proposed algorithm is applied on different remote sensing images taken on surface of the earth for identifying minerals. Here we get concentrated on giacomo mineral which appears to be in white color. It exists just below the surface of the earth. Basing on the spectral appearance from an image we proposed the algorithm containing k-means clustering technique and edge detection. The resulting images show the presence of the mineral Giacomo in different parts of the world.

Sobel edge detection method is considered in this algorithm to extract the edges of the area where the mineral is located. The common Sobel edge detector which have 3×3 horizontal and vertical masks is used in the edge function, in the image toolbox of matlab. The whole algorithm was programmed and implemented in MATLAB. Here spectral appearance is considered as a parameter to extract the minerals in the image. Similarly we can extend the same process for detecting other minerals in future development. The main disadvantage of this process is that it detects the minerals existing just below the surface of the earth.

REFERENCES

[1] The history of k-means type of algorithms (LBG Algorithm, 1980) R.M. Gray and D.L. Neuhoff, "Quantization," *IEEE Transactions onInformation Theory*, Vol. 44, pp. 2325-2384, October 1998. (Commemorative Issue, 1948-1998)

[2] An Introduction to Data Mining, Tan, Steinbach, Kumar, Addision-Wesley, 2005. http://www-users.cs.umn.edu/~kumar/dmbook/index.php

[3] SOBEL, I., *An Isotropic 3×3 Gradient Operator,* Machine Vision for Three – Dimensional

[4] SOBEL, I., Camera Models and Perception, Ph.D. thesis, Stanford University, Stanford,CA, 1970..

[5] Data Mining: Concepts and Techniques, 2nd Edition, Jiawei Han and Micheline Kamber, Morgan Kauffman, 2006 http://www-sal.cs.uiuc.edu/~hanj/bk2

[6] K-means tutorial slides (Andrew Moore) http://www.autonlab.org/tutorials/kmeans11.pdf

[7] ISPRS Journal of Photogrammetry and Remote Sensing http://www.asprs.org/

[8] http://csep10.phys.utk.edu/astr162/lect/ligh t/ radiation.html