OPTIMAL PROJECTION FROM N-DIMENSIONAL PATTERN SPACE INTO A PLANE

Fouad N. Hassan
University of Baghdad, College of Science, Department of Astronomy.

Abstract

This paper presents a discriminant algorithm that seeks to separate different classes as much as possible for discriminant analysis or dimension reduction. The optimization is achieved through the maximization of the Fisher ratio (which is defined as the ratio of the between-class scatter to the sum of within-class scatters).

This algorithm for feature extraction shows improvement over the conventional feature selection algorithms used in remote sensing as well with other applications. The conducted experiments are accomplished using both simulated Gaussian and real airborne MSS/TM satellite data for both large and small sample size. Although the conducted experiments are performed over the case of two classes, extension to n-dimensions can be easily obtained using the binary decision tree.

Introduction

Feature extraction plays important role in the problems of pattern classification. By feature extraction the n-dimensional space is reduced to a lower dimensional one. This reduction is performed under the condition that certain criteria are preserved or minimized.

This problem is special importance in the classification of remotely sensed data. This follows from the fact that the number of the training samples (i.e., pre-labeled samples) are usually difficult or expensive to obtain. Furthermore, the number of the bands in the multi spectral scanner may be as large as 24and this number is expected to be increased to 50-100 in the future [1]. This increase in the number of dimensions with the limited sample size will lead to Hughes phenomena [2]. This causes the increase probability of classification error with the increase of the number of bands. Consequently, dimensionality reduction has to be performed to improve classification accuracy.

Feature extraction or selection in remote sensing is usually performed using the KL transform [6], the divergence measure or the J.M distance. In this paper, feature extraction is accomplished by minimizing the sum of the within class scatters and maximizing the between-class scatter.

Fisher Dimensionality Reduction

Consider the two sets of samples A and B. Let A contains a samples and B contains b samples. Each sample in the two classes A and B is represented by n-dimensional vector.

Let x_i and y_j ($1<i<a$, $1<j<b$) be the vector in A and B respectively. The means m_1 and m_2 of A and B are given by:

$$m_1 = \frac{1}{a}\sum_{i=1}^{a} x_i \quad \text{and} \quad \mathrm{m}_2 = \frac{1}{b}\sum_{j=1}^{b} y_j \quad \text{............} (1)$$

Therefore, the distance between the two classes is given by:

$$D = m_1 - m_2$$

The within class scatters W_1 and W_2 of A and B are given by:

$$W_1 = \sum_{i=1}^{a} (x_i - m_1)^t (x_i - m_1), \quad \text{..................} (2a)$$

And,

$$W_2 = \sum_{j=1}^{b} (y_j - m_2)^t (y_j - m_2), \quad \text{................} (2b)$$

Where, t denotes the transpose of the matrix. The total within class matrix is equal to $W_1+W_2=W$.

Let d_1, d_2,….., d_m be m orthonormal (1x m) vectors (m<n)over which the projection is performed. These vectors are used to project the samples x_i and y_j from the original space into m-dimensional space to get the samples u_i and v_j respectively. The samples *u* and *v* are (1xm) vectors that can be written in the form

$$u_i^t = (u_{1i}, ..., u_{ki}, ..., u_{mi})$$
$$v_j^t = (v_{1j}, ..., v_{kj}, ..., v_{mj})$$

Where

$$u_{ki} = d_k x_i \quad and \quad v_{kj} = d_k y_j \text{ ....(3)}$$

In the m-dimensional space the between class scatter N is given by:

$$N = (d_1^t D)^2 + (d_2^t D)^2 + ... + (d_m^t D)^2$$
$$= \sum_{i=1}^{m} (d_i^t D)^2 \text{ ....(4)}$$

While the total within class scatter *D* is given by

$$D = \sum_{i=1}^{m} d_i^t W d_i \quad ,\text{…………………….....(5)}$$

In this paper the projection is performed through the maximization of the ratio (F) of the between class scatter to the within class scatter, i.e.,

$$Maximize \quad F = \frac{N}{D}$$

Subject to the condition that the vectors d_1, d_2,….,d_m are orthonormal.

The projection over a plane is of significant importance. This follows from the fact that the intrinsic dimensionality of the available remote sensing date is two. Furthermore the projection over plane can be used in the on-line interactive graphic systems [3]. For the projection over plane *F* is given by:

$$F = \frac{(d_1^t D)^2 + (d_2^t D)^2}{d_1^t W d_1 + d_2^t W d_2} \text{ ....(6)}$$

The vectors d_1 and d_2 that maximize the above ratio (subjected to the condition that d_1 and d_2 are orthonormal) can be evaluated by a simple iterative method. This method starts with two orthonormal vectors $d_1(1)$ and $d_2(1)$. For each iteration (*i*), one of the vectors, say $d_1(i)$, is kept constant while the other $d_2(i)$ is evaluated to maximize the objective function:

$$\begin{aligned} F(i+1) = &\Big((d_1^t(i+1)D)^2 + (d_2^t(i+1)D)^2\Big) / \\ &\Big(d_1^t(i+1)Wd_1(i+1) + d_2^t(i+1)Wd_2(i+1)\Big) \\ &+ l_1[d_1^t(i+1)d_2(i+1)] \\ &+ l_2[d_2^t(i+1)d_2(i+1) - 1] \end{aligned} \quad ,\text{..................... (7)}$$

Where, l_1 and l_2 are the Lagrange multipliers.

For each iteration the vector d_1 is kept constant, i.e.

$$d_1(i+1) = d_1(i) \text{ .... (8)}$$

Using Newton method, the vector $d_2(i+1)$ can be written as:

$$d_2(i+1) = d_2(i) + d\, d_2(i) \text{ .... (9)}$$

Where

$$d\, d_2(i) = e\, \nabla_{d_2(i)} F(i) \text{ ....(10)}$$

The constant ε determines the speed by which the iterative process will converge. Also, it determines the accuracy in evaluating the two vectors d_1 and d_2. Too small values for ε will decrease the speed of convergence while high values for ε will decrease the accuracy and increase the speed of convergence. The vector $\nabla_{d_2(i)} F(i)$ is obtained by differentiating *F(i)* with respect to the vector $d_2(i)$ to get

$$\begin{aligned} \nabla_{d_2(i)} F(i) = &\Big(2D(i)\big[d_2^t(i)D\big]^2 - 2N(i)Wd_2(i)\Big) / \\ &\big[D(i)\big]^2 + l_1 d_1(i) + l_2 d_2(i) \end{aligned} \quad ,\text{..............(11)}$$

Where,

$$D(i) = d_1^t(i)Wd_i(i) + d_2^t(i)Wd_2$$

and

$$N(i) = [d_1^t(i)D]^2 + [d_2^t(i)D]^2$$

To evaluate $\nabla_{d_2(i)}$, the values of l_1, l_2 have to determined. Since

$$d_1^t(i+1)d_2(i+1)=0 \text{ .... (12a)}$$

and,

$$d_2^t(i+1)d_2(i+1)=1 \text{ .... (12b)}$$

Using equation's (8) and (9), one will get:

$$d_1^t(i)d\,d_2(i)=0 \text{ .... (13a)}$$

and,

$$d_2^t(i)d\,d_2(i)=0 \text{ .... (13b)}$$

Multiplying equation (11) by $d_1^t(i)$ and using equation (13a), l_2 can be determined. Also, multiplying equation (11) by $d_2^t(i)$ and using equation (13b), l_1 can be determined. The vector $\nabla_{d_2(i)}F(i)$ is then evaluated using equation (11). Furthermore, the vector $d_2^t(i+1)$ can be evaluated (for the next trial *i+l*) using equation (9). The two vectors derived in Ref [3] can be used for the first trial, where

$$d_1(1)=a\,W^{-1}(m_1-m_2) \text{ ,........................... (14)}$$

and

$$d_2(1)=b\left\{W^{-1}-\frac{\Delta^t\left[W^{-1}\right]^2\Delta}{\Delta^t\left[W^{-1}\right]^3\Delta}\left[W^{-1}\right]^2\right\}\Delta \text{ ,...... (15)}$$

Where, a and b are the normalization constants.

Comparison between feature extraction methods

In this section, it will be shown that the Fisher feature extraction possesses some interesting properties that make it superior over other methods especially in remote sensing problems.

In the following, it will be proved that for any projection over a plane, the probability of classification error will have a lower limit. This limit is maximized by the projection over Fisher plane. This result is achieved through extending the Chebyshev inequality.

Let *f(x,y)* and *g(x,y)* be the probability density functions of classes 1 and 2, respectively, after projection over a plane by any of the feature extraction methods. In this plane, let h_1, h_2 be the distance between the means of the two classes in the first and second dimensions. Also, let S_1 and S_2 be the within class scatters of classes 1 and 2 respectively. Then

$$E(x^2+y^2)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x^2+y^2)f(x,y)dxdy \text{ ,.............................. (15)}$$

Since $E(x^2+y^2)$ is equal to S_1, then from the above equation one may get:

$$S_1>\int_{|x|>(h_1/2)}\int_{|y|>(h_2/2)}(x^2+y^2)f(x,y)dxdy$$

or,

$$S_1>\left[\left(\frac{h_1}{2}\right)^2+\left(\frac{h_2}{2}\right)^2\right]\int_{|x|>(h_1/2)}\int_{|y|>(h_2/2)}(x^2+y^2)f(x,y)dxdy$$

But the above integral is equal to the probability of classification error in class 1 (P_1). Thus,

$$4S_1>(h_1^2+h_2^2)P_1 \text{ .... (16)}$$

Similarly, for class 2:

$$4S_2>(h_1^2+h_2^2)P_2 \text{ ,.................................... (17)}$$

Where, P_2 is the probability of classification error in class 2.

Summing inequalities (16) and (17) to get

$$4(S_1+S_2)>(h_1^2+h_2^2)(P_1+P_2) \text{ ,............... (18)}$$

But (S_1+S_2) is equal to the total within class scatter and $(h_1^2+h_2^2)$ is equal to the between class scatter. Furthermore, (P_1+P_2) is equal to the total probability of error (P_e). Consequently, equation (18) is reduced to

$$P_e<\frac{4}{F} \text{ ,.. (19)}$$

It should be noted that the above inequality doesn't work for the projection over a plane only. It can be applied for the projection over any space of dimensions lower than the original space. It is obvious that equation (19) determines a lower limit for the probability of error. But the aim of Fisher dimensionality reduction is to maximize F. Thus by this projection, the lower limit is maximized.

It has been stated [4] ,[7] that for the multi-variant Gaussian case (with $U_k(i)$ and s_k denoting, respectively, the mean and variance of k in class i) with variances assumed equal for both classes, then if the variables are independent and

$$g_k = \left| \frac{U_k(1) - U_k(2)}{s_k} \right| > 0 \quad for\ all\ k$$

then the probability of error tends to zero with increasing the number of the dimensions if $\sum_{k=1}^{\infty} g_k$ diverges.

In the following it will be proved that the above result can be obtained for any probability density function after projection over Fisher discriminate vector. In another word, after projection over Fisher discriminate vector, the probability of error tends to zero when

$\sum_{k=1}^{\infty} g_k$ diverges.

This result follows from equation (19) (where F is evaluated after projection over the vector $(a_1,\ldots,a_i,\ldots,a_n)$). Then:

$$F = \left(\sum_{i=1}^{n} a_i D_i \right)^2 \Bigg/ \left(\sum_{i=1}^{n} a_i^2 s_i^2 \right) \quad ,............. (20a)$$

Where,

$$\sum_{i=1}^{n} a_i^2 = 1 \quad ,... (20b)$$

Choosing $a_i=\alpha(\Delta/\sigma_i^2)$, where α is the normalization constant, then equation (20a) becomes:

$$F = \left(\sum_{i=1}^{n} a_i D_i \right)^2 \Bigg/ \left(\sum_{i=1}^{n} a_i^2 s_i^2 \right)$$

This ratio [8] tends to ∞ as n approaches ∞, since, by the projection over the Fisher discriminant vector ***F*** is maximized, then by this projection ***F*** should tend to ∞. Consequently, by the virtue of equation (19) the probability of error tends to zero.

From the above result, it may be concluded that adding more informative dimensions will improve the classifier performance. This important result is not guaranteed when using KL transform. On the contrary, adding more dimensions in the KL transform may spoil the whole performance. This follows from the fact that this transform gives projection over the dimensions of higher variance. However, the variance in the data may come from the within or between class scatters. Since the KL can' t discriminate between these two scatters; then adding dimensions of high within class scatters will deteriorate the classifier performance.

By the feature selection methods such as divergence and J.M distance, a subset of dimensions is selected for the classifier design. Therefore a considerable loss in the classification accuracy will occur if the original space contains many valuable dimensions.

Experimental results

Simulated Gaussian (using central point theorem) and MSS data have been used for the comparison between feature extraction methods.

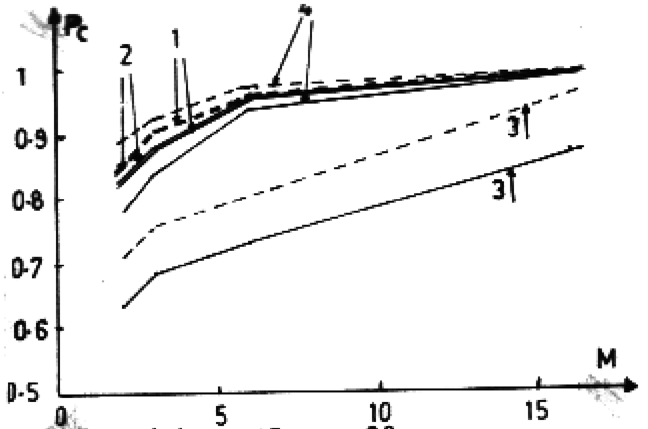
1. Simulated results: sets of two n-dimensional Gaussian classes are generated for the test. For each test, the two classes have the same covariance matrix and each contains (a) samples. These samples are called the design set that can be used for the classifier design. The generator, that is used to generate the design set, is also used to generate 200 test samples. The test samples are then classified by the classifier (designed by the design set). Four methods of classification are used with each set. The first one uses the Bays classifier over the original n-dimensional space. The second one is performed through the projection over the Fisher plane and applying the Bays rule using two dimensional data. The third and forth methods are performed by

applying the Bays rule with two dimensional data. The data is obtained by the projection over a plane using the KL and the divergence methods respectively. Graphs (1) to (8) show samples of the results. For each graph, (n) and (a) are kept constant and the classifier performance is evaluated for different values of the Mahalanobis distance $M=(m_1- m_2)^t \sum^{-1}(m_1- m_2)$. Form this distance, the minimum probability of error P_e can be evaluated where $P_e=0.5-erf(M^{1/2}/2)$. Graphs (1) to (4) show the results for the date with equal (D / s) for all the dimensions. From these graphs it is obvious that, with the divergence methods, there is a considerable loss in the classifier accuracy compared with the other methods. Graphs (5) to (8) shows the results for different (Δ/σ). From these graphs it is clear that the KL method doesn’t give acceptable results compared with the other methods. From graphs (1) to (8) it is clear that, with the Fisher method, the classifier performance is improved with increasing the number of the dimensions. This result is not guaranteed when using other methods.

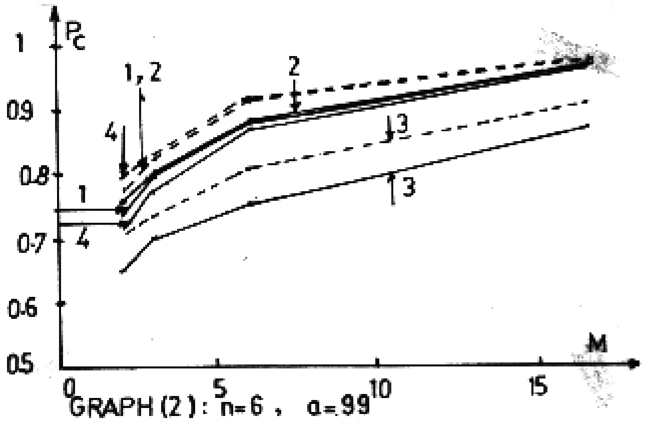
2. TM/MSS data: the feature extraction methods mentioned previously have been used to discriminate between the water and vegetation obtained from 6 bands TM/MSS data. Fifteen samples for each class are used in the test. With this low sample size the U method or "leave-one-out" is recommended [5] to get good results. By this method, 14 samples were used as the design set, while the remaining one is used as the test set. The test is repeated 15 times and the average of results is evaluated. With the Fisher method, it is found that the probability of error is equal to 3.3% for the test set while it is equal to 3.6% for the design set. For the KL method the probabilities of error are equal to 14.3% and 16.80% respectively. For the divergence methods, these are equal to 14.3% and 9% respectively. By using the Bays rule in the original space, these probabilities are equal to 14.3% and 6% respectively. From as these result, it may be concluded the Fisher method gives the best results.

Conclusions

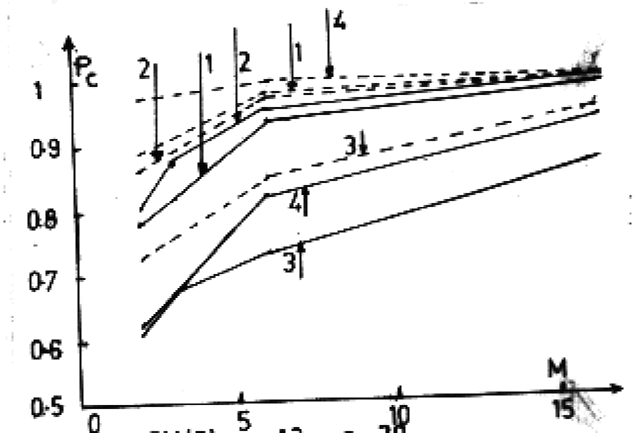
Feature extraction method by maximizing the Fisher ratio is presented. In this paper the emphasis is on the projection over a plane. It has been found that with increasing the number of the dimensions, the Fisher method will give better performance. This result is not guaranteed when using the KL method. Furthermore, there may be a considerable loss in the classification accuracy when using any of the feature selection method. Thus the Fisher method is more suitable because of the large number of available bands in remotely sensed date and this number is expected to be increased in the future.



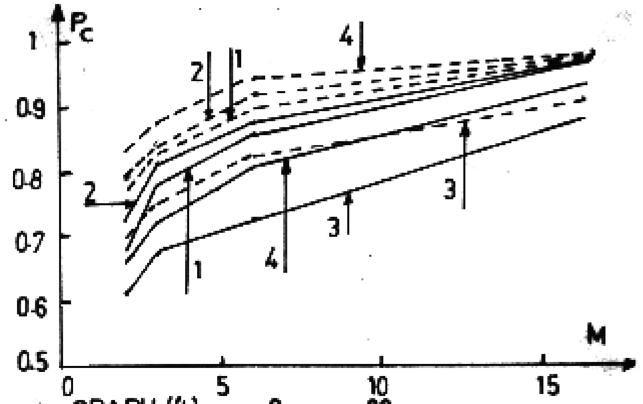
GRAPH (1): n=12 , a=99



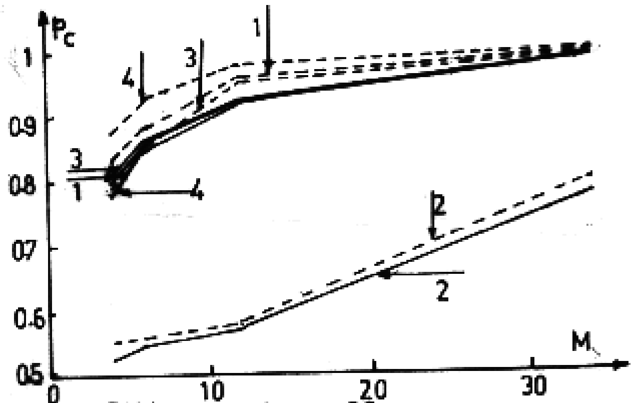
GRAPH (2): n=6 , a=99



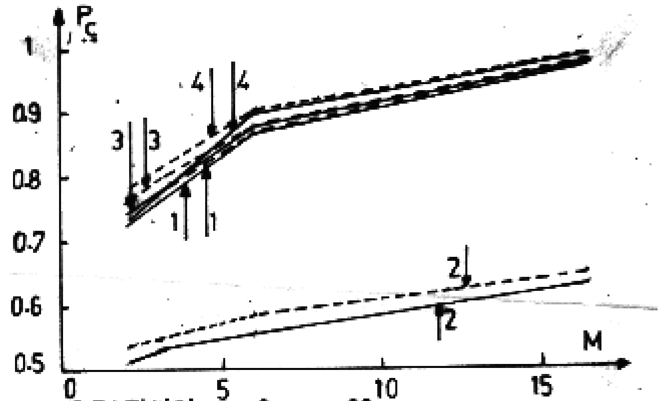
GRAPH (3): n=12 a=20



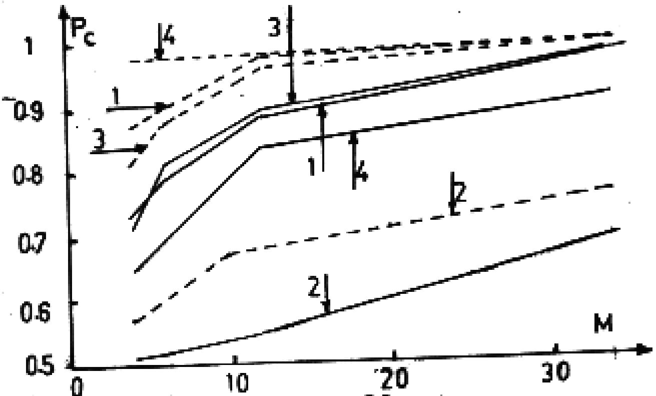
GRAPH (4): n=6 , a=99



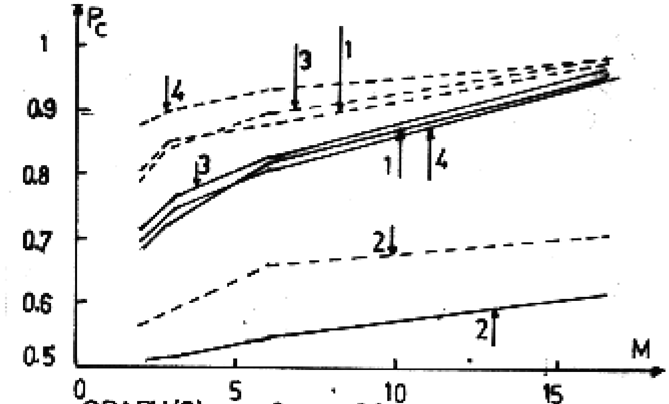
GRAPH (5): n=12 , a=99



GRAPH (6): n=6 , a=99



GRAPH(7) : n=12 , a=20



GRAPH (8) : n=6 , a=20

Graphs 1 to 8 represent probabilities of correct classification vs. Mahalaobis distance for sample sizes (α=99 and 20) and dimensionalities (n=12 and 6). In each graph, (1) refers to Fisher method, (2) refers to KL method, (3) refers to divergence method and (4) refers to Bays method over n-dotted curves refer to tests over design sets while continuous curves refer to tests over test sets.

References

[1] M.J. Muasher and D.A. Landgrebe, "A binary tree feature selection technique for limited training sample size", Remote Sensing of Environment, vol. 16, no.3, pp.183-194, Dec.,1984.

[2] G.H. Hughes, "On the mean accuracy of statistical pattern recognizer", IEEE Trans. on Inform. Theory, vol. 14, pp. 55-63, Jan., 1968.

[3] J.W. Sammon (JR), "An optimal discriminate plane", IEEE Trans. on Comput., vol.19, pp. 826-829, Sept., 1970.

[4] L. Kanal and B. Handraskaran, "On dimensionality and sample size in statistical pattern classification", Proc. Nat. Electron. Conf., vol. 24, pp. 2-7, Dec., 1968.

[5] G.T. Toussaint, "Bibliography on estimation of misclassification", IEEE trans. on Inform. Theory, vol. 20, pp. 472-479, July, 1974.

[6] R.O. Duda, P.E. Hart, and D.G. Stock, “Pattern Classification”, 2nd Edition, Wiley-interscience, Chichester (2000).

[7] P.N. Belhumeur, J.P. Hespanha, and D.J. Kriegman, “Eigen faces vs. Fisher faces: Recognition using class specific linear projection”, IEEE Transactions on Pattern Analysis and Machine Intelligence, v.19, n.7 , pp. 711-720, July 1997.

[8] J. Lu, K.N. Plataniotis, A.N. Venetsanapoulos ,”Face recognition using kernel direct discriminant analysis algorithms”. IEEE Transactions on Neural Networks v. 14, no. 1, pp. 117-126, 2003.

الخلاصة

هذا البحث يعرض خوارزمية للتمييز تؤدي الى فصل الاصناف المختلفة القابلة لتحليل المميز او اختزال البعد .ان الخوارزمية المثلى نفذت من خلال نسبة فشرالعظمى (والتي تعرف على انها النسبة بين الانتشار بين الاصناف الى مجموع الانتشار ضمن الصنف).وأن هذا الخوارزمية قامت بأستخلاص الملامح مما يؤدي الى تحسين فصل الاصناف مقارنة الى خوارزميات اخرى بأختيار الخصائص التقليدية المستخدمة في التحسس النائي اضافة الى التطبيقات الاخرى .

أن التجارب المستخدمة قد نفذت باستخدام كلا من البيانات الممثلة بكاوس والبيانات الحقيقية (MSS,TM) عبر الاقمار الصناعيةلأغراض مسح الموارد الطبيعية ولعينات كبيرة وصغيرة الحجم.

أن التجارب المستخدمة التي نفذت لصنفين تبين انه يمكن تطبيقها على الفضاء ذو البعد-ن وذلك من خلال أستخدام شجرة القرار النهائي (binary decision tree).