

Design and Implementation of Classification System Based On Soft Computing and Statistical Approach

Subodh Prasad, Dhyan Singh Rawat

Amrapali Institute of Technology & Sciences, Uttarakhand, India

Abstract— A classifier based on statistical methods and soft computing approaches can be capable of identifying the mines and non mines. In this paper the design and development of such classifier has been discussed using the various clustering, classification and rules establishment algorithms and finally the algorithm has been compared on the basis of complexity and accuracy. Designing such a classifier is a big challenge as the data is not linearly separable and have overlapping features, so it's not possible to design such a classifier with 100% accuracy.

The project deals with PVC tubes, wood pieces and copper cylinders as non-mine data in addition to data of various mines. The basic idea of the classification is based on the fact that it's safer to predict the non-mines data as mines, than to predict mines data as non-mines. The unsupervised learning based ART algorithm divides the data into several clusters which are merged on the basis of above fact. Genetic algorithm has enhanced the results, to establish the results having negation in the antecedent part. Fuzzy approaches has been used to give the membership values corresponding to each class in order to visualize the class of data in a better way.

Keywords- Algorithm; Fuzzy Logic, Clustering

INTRODUCTION

"If one already know about the upcoming hazards; it is very easy to find the way to abolish it."

Here, this sentence is being described in the context of Landmine Detection and Decontamination. The main objective is to predict whether particular point of working envelope is occupied by mines or not with some confidence parameter. Robot is designed to move toward these predicted areas to decontaminate the mines. These mines occupied area can be known before initiation of robot movements or can be predicted dynamically. To design an obstacles free path for robot is the another aspect.

To tackle this problem a classification toolkit has been designed using some statistical and soft computing based approaches to cluster the data, to predict the possible class of incoming data, to generate some rules in the term of confidence parameter. The data may be given in image form or some tabular form having all numeric or categorized attributes.

It is impossible to design a classifier having 100% right classification because it is not easy to differentiate between the data of metallic debris, PVC tubes and actual mine data. On the

basis of this prediction path designers develop the obstacle free path to decontaminate these mines.

Anti-Personal landmines are a significant barrier to economic and social development in a number of countries. So a classification system is needed that can differentiate a mine from metallic debris on the basis of given data. This data is generated by some highly accurate sensors.

CHALLENGES IN THIS FIELD

In the field of classification and rules establishment, the basic problems are the features extraction (building blocks of algorithms) and selection of best algorithms those can generate results with high certainty value.

FEATURE EXTRACTION

The initial problem is the problem of features extraction. Generally the image data is given having a blurred image of an object so it is very difficult to extract the exact boundary of the object. There may be various features which can be used as the raw material of system. Here blob size, blob-aspect-ratio and blob-intensity have been chosen.

The given data may contain the images of PVC tube, metallic debris and Mines. The data in some tabular format having numerical or categorized values of attributes can also be processed, which is more suitable for the algorithms.

SELECTION OF AN ALGORITHM

The second problem is to choose an algorithm that can interpret the problem in best way. The algorithms can be categorized in two parts:

1. Statistical approaches
2. Soft computing based approaches

The three types of algorithms can be applied here: Classification, Clustering and Rules establishment with some certainty factor. The best way is to design various algorithms and then check their efficiency and accuracy.

APPROACHES IN THIS DIRECTION

In this section the various algorithms will be discussed which are to be used to achieve the objective.

4A. Statistical Approaches

Two categories of algorithms has been used, one for clustering (K-mean algorithm) and another for the prediction of the class of incoming data (K nearest neighbor).

4A.Clustering Algorithms:

4A.1. K-mean

Clustering is a nonlinear activity that generates ideas, images and feelings around a stimulus word. Clustering may be a class or an individual activity.

If the number of data is less than the number of clusters then each data point is assigned as the centroid of the cluster. Each centroid is identified by a cluster number. If the number of data is bigger than the number of cluster, then for each data point the distances between all centroids are calculated and minimum of them is selected. The incoming data-point is assigned a cluster which has minimum distance. Since the location of the centroid is not sure, one need to adjust the centroid location based on the current updated data. Then all the data has been assigned to this new centroid. The process is being repeated until no data has been moving to another cluster anymore. Mathematically this loop can be proved to be convergent

Since the problem has only two classes mine and non-mine, so the number of classes is given 2 as input with the dataset [B1].

4A.2. K-nearest Neighbor

Nearest neighbor technique is used to predict the class of incoming data on the basis of given training data and density estimator (k-NN) to estimate the confidence of the incoming sample for a particular class. Finally a class is predicted having the highest estimator.

Density estimator: $qc(x) = (\text{number of neighbors of class } c)/K$

The neighbors are the k closest point to the given sample .Their mutual distances are calculated by city block distance [B2]. The problem of choosing k still remains, but a general rule of thumb can be used

$K = \sqrt{N}$.

where N is the number of learning samples.

The main disadvantage of this method is that it is computationally intensive for large data sets.

4B. Soft computing approaches

Soft computing approaches can be classified into several categories like Neural approaches, Fuzzy clustering, Adaptive resonance theory, Kohonen SOM, Genetic algorithm etc.

4B.1.Genetic algorithm to establish rules

To establish the rules between the attributes of data, association rule has been used but association rule mining cannot predict the complete set of rules, i.e. the rules which have negation in the attributes cannot be discovered. To overcome this disadvantage, Genetic Algorithms (GAs) has been used.

First of all, association rule is applied with some support and confidence values entered by user to generate some base rules and these rules are sent to genetic algorithm as input which helps to evolve some new rules having negation in attributes.

The three basic part of genetic algorithm is as follow:

- (a) Selection: Roulette wheel technique is used to select the two parents[R1].
- (b) Crossover: A random point (crossover point) is generated and the segment to the left of this point of first parent and that of second parent are interchanged.
- (c) Mutation: Mutation point is generated randomly and the bit value at this point has been toggled.

After some iteration one can find some rules following the above properties and having high fitness value that can be calculated either using the confidence value or by confusion matrix.

4B.2.Adaptive resonance theory (ART):

Back propagation network is very powerful and it can simulate any continuous function given a certain number of hidden neurons and a certain form of activation functions. But once the back propagation is trained, the number of hidden neurons and the weights are fixed. The network cannot learn from the new patterns unless the network is re-trained from scratch, so there is no plasticity [R2].

So ART, a new neural network technique has been used to solve this problem. In it final objective is to cluster the data in several chunks.

One by one, samples from the data as input neurons is sent as input and the activation value is calculated corresponding to each of the existing output neurons and the highest value is chosen, if this value is higher than threshold value then the weight of this connection is updated otherwise a new output neuron has to be added. After certain iterations one can find the proper clusters of the data. In this application, only two classes are there (mine and non-mine). Another fact is that it's safer to predict the non-mines data as mine, than to predict mines data as non-mines because it may be dangerous. Among all the clusters, the cluster having the cluster center farthest from the mine data center is been classified as nonmine, and rest of the clusters are classified as mine.

Here activation function is calculated as the city block distance of the incoming normalized data and weights of connection.

4B.3 Fuzzy c mean:

In the classical clustering algorithm there is a crisp membership of a class (either one or zero) but while classifying the mine-data it's not easy to differentiate between mine and non-mine. So method has to be searched which can tell the membership of the data in each class. If this membership is average then we have to deal this data as special data and classify this in class of mine (as mine are dangerous !) [R3].

$\forall j \in \{1, 2, \dots, |X|\}$ where $|X|$ is the feature vector

$$\sum_{i=1}^{|P|} u_{i,j} = 1 \quad \text{and } p \text{ is the number of classes (p=2 in our case)}$$

Membership values

$$u_{i,j} = \frac{1}{\sum_{k=1}^{|P|} \left(\frac{d_{i,j}^2}{d_{k,j}^2} \right)^{\frac{1}{m-1}}}$$

Euclidean distance :

$$d_{i,j}^2 = \|\vec{x}_j - \vec{c}_i\|^2$$

Mean center prototype :

$$\vec{c}_i = \frac{\sum_{j=1}^{|X|} u_{i,j}^m w_j \vec{x}_j}{\sum_{j=1}^{|X|} u_{i,j}^m w_j}$$

Mean center prototype(Ci)=

$$\vec{c}_i = \frac{\sum_{j=1}^{|X|} u_{i,j}^m w_j \vec{x}_j}{\sum_{j=1}^{|X|} u_{i,j}^m w_j}$$

If the difference of the membership value with previous membership value is less than the threshold, then the algorithm terminates having the membership value for each class.

4B.4. Gustavson-Kessel Algorithm

It's an improvement of fuzzy c mean clustering algorithm. The correlation between the data is not considered in c mean, in this algorithm distance formula is redefined as: [R3]

Mahalanobis distance:

$$d_{i,j}^2 = (x_j - c_i)^T A_i (x_j - c_i)$$

where A_i is the mean center prototype and x_j and c_j are the sample attribute and cluster center.

Covariance matrix is calculated as :

$$S_i = \frac{\sum_{j=1}^{|X|} u_{i,j}^m w_j (x_j - v_i)(x_j - v_i)^T}{\sum_{j=1}^{|X|} u_{i,j}^m w_j}$$

Mean center prototype

$$A_i = \frac{1}{\det(S_i)} S_i^{-1}$$

4B.5. Gath-Geva Algorithm :

This algorithm assumes that data is normally distributed [R3].

Distance can be calculated as

$$d_{i,j}^2 = \frac{\sqrt{\det(S_i)}}{P_i} \exp \left(\frac{1}{2} (x_j - c_i)^T A_i (x_j - c_i) \right)$$

where

$$P_i = \frac{\sum_{j=1}^{|X|} u_{i,j}^m w_j}{\sum_{j=1}^{|X|} \sum_{t=1}^{|P|} u_{t,j}^m w_j}$$

is the a-priori probability of data

belonging to cluster i , and $A_i = S_i^{-1}$ is the mean center prototype

Before applying this algorithm it is suggested to analyze data whether it is normally distributed or not.

4B.6. Kohonen SOM:

A competitive network learns to categorize the input vectors presented to it. If a neural network only needs to learn to categorize its input vectors, then a competitive network will do. Competitive networks also learn the distribution of inputs by dedicating more neurons to the classifying parts of the input space with higher densities of input [B3].

A self-organizing map learns to categorize input vectors. It also learns the distribution of input vectors. Feature maps allocate more neurons to recognize parts of the input space where many input vectors occur and allocate fewer neurons to the parts of the input space where very few input vectors occur.

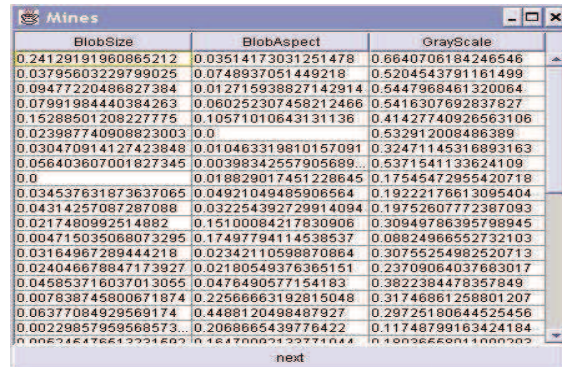
Self-organizing maps also learn the topology of their input vectors. Neurons next to each other in the network learn to respond to similar vectors. The layers of neurons can be imagined to be a rubber net that is stretched over the regions in the input space where input vectors occur. Self-organizing maps allow neurons that are neighbors to the winning neuron to output values. Thus the transition of output vectors is much smoother than that obtained with competitive layers, where only one neuron has an output at a time.

RESULTS

As it's already been discussed that input can be in image form or tabular form. Matlab tool or any other image processing tool can be used to extract the features from the input images. The numerical attributes based table is also required with the

entry whether the data belongs to mine or non-mine. For the genetic algorithm, categorized table is needed. For this purpose data has been categorized in three categories: low, medium, high with class value mine or non-mine.

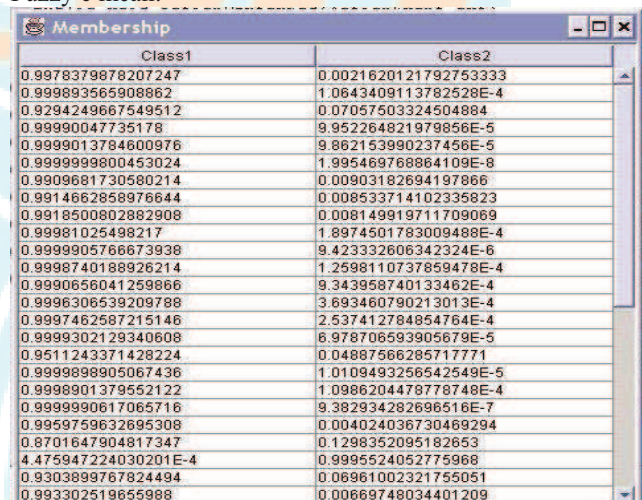
The objective of this comparison based approach is to compare between various algorithms and differentiate them on the basis of their accuracy and results.



BlobSize	BlobAspect	GrayScale
0.24129191960865212	0.03514173031251478	0.6640706184246546
0.03795603229799025	0.0748937051449218	0.5204543791161499
0.09477220488827384	0.012715938827142914	0.5447968481320084
0.07991984440384263	0.060252307458212466	0.5416307692837827
0.15288501208227775	0.10571010643131136	0.41427740926563106
0.023987740908823003	0.0	0.532912008486389
0.030470914127423848	0.010463319810157091	0.32471145316893163
0.056403607001827345	0.00398342557905689...	0.5371541133624109
0.0	0.018829017451228645	0.17545472955420718
0.034537631873637065	0.0492104948506564	0.19222176613095404
0.04314257087287088	0.032254392729914094	0.19752607772387093
0.0217480992514882	0.1510084217830908	0.30949786395798945
0.004715035088073295	0.17497794114538537	0.08824966552732103
0.03164967289444218	0.02342110598870864	0.30755254982520713
0.024046678847173927	0.02180949376365151	0.23709064037683017
0.045853716037013055	0.0476490577154183	0.3822384478367949
0.007838745800671874	0.22566663192815048	0.31746861258801207
0.06377084929569174	0.4488120498487927	0.29725180644525456
0.00229857959568573...	0.2068665439776422	0.11748799163424184
0.002516476612231592	0.1647002132771044	0.18036559011000202

The ART gives the multiple class distribution of the given data. To predict a non-mine as a mine is not as much dangerous as to predict a mine as non-mine, so all the clusters having more distance from the non-mine center can be assigned mine class. Thus merging of data is boosting up this algorithm.

Fuzzy c mean:



Class1	Class2
0.9978379878207247	0.002162012179275333
0.999893565908862	1.0643409113782528E-4
0.9294249667549512	0.07057503324504884
0.99990047735178	9.952264821979856E-5
0.9999013784600976	9.862153990237456E-5
0.999999800453024	1.995469768864109E-8
0.9909681730580214	0.00903182694197866
0.9914662858976644	0.008533714102335823
0.9918500802882908	0.008149919711709069
0.99981025498217	1.8974501783009488E-4
0.9999905766673938	9.42332606342324E-6
0.9998740188926214	1.2598110737859478E-4
0.9990656041259866	9.343958740133462E-4
0.9996306539209788	3.693460790213013E-4
0.9997462587215146	2.537412784854764E-4
0.9999302129340608	6.978706593905679E-5
0.9511243371428224	0.04887566285717771
0.999898905067436	1.0109493256542549E-5
0.9998901379552122	1.0986204478778748E-4
0.9999990617065716	9.382934282696516E-7
0.9959759632695308	0.004024036730469294
0.8701647904817347	0.1298352095182653
4.475947224030201E-4	0.9995524052775968
0.9303899767824494	0.06661002321755051
0.993302519655988	0.00669748034401209

The fuzzy c mean algorithm is giving rules with membership value in each class. Now it's very easy to check some data that cannot be classified as mine and non-mine, so this type of data can be put into mine class to avoid danger.

Genetic algorithm:

```
Association::

(BlobAspect is low ) -->not mine with confidence 0.5
(GrayScale is low ) --> mine with confidence 0.4318181818181818
(BlobSize is low ) --> mine with confidence 0.36363636363636365
(BlobAspect is high ) --> mine with confidence 0.3409090909090909
(GrayScale is high ) -->not mine with confidence 0.29545454545454547
(BlobAspect is high ) and(GrayScale is low ) --> mine with confidence 0.29545454545454547
(BlobAspect is low ) and(GrayScale is high ) -->not mine with confidence 0.29545454545454547
(BlobSize is low ) and(GrayScale is low ) --> mine with confidence 0.29545454545454547
(BlobSize is low ) -->not mine with confidence 0.2727272727272727
(BlobSize is low ) and(BlobAspect is low ) -->not mine with confidence 0.2727272727272727Genetic rule::

(BlobSize is not low ) and(BlobAspect is high ) and(GrayScale is high ) --> mine with confidence 1.0
(BlobSize is not low ) and(BlobAspect is high ) and(GrayScale is high ) --> mine with confidence 1.0
(BlobSize is not low ) --> mine with confidence 0.625
(BlobSize is not low ) --> mine with confidence 0.625
(BlobSize is low ) and(BlobAspect is not high ) --> mine with confidence 0.6666666666666666
(BlobSize is not low ) --> mine with confidence 0.625
(BlobSize is not low ) and(BlobAspect is high ) and(GrayScale is high ) --> mine with confidence 1.0
(BlobSize is not low ) --> mine with confidence 0.625
(BlobSize is not low ) --> mine with confidence 0.625
(BlobSize is not low ) --> mine with confidence 0.625
```

This snapshot is displaying the result of both association rule as well as genetic rules. It's very much clear that genetic algorithm has generated the rule having negative attribute value in antecedent part so this algorithm is very useful to establish rules.

ART:

Kmean Algorithm:

BlobSize	BlobAspect	GrayScale
0.039213374982809815	0.2812782681680574	0.27455502977348356
0.09477220486827384	0.012715938827142914	0.5447968461320064
0.07991984440384263	0.060252307458212466	0.5416307692837827
0.15288501208227775	0.10571010643131136	0.41427740926563106
0.023987740908823003	0.0	0.532912008486389
0.030470914127423848	0.010463319810157091	0.32471145316893163
0.056403607001827345	0.0039834255790568915	0.5371541133624109
0.0	0.018829017451228645	0.17545472955420718
0.034537631873637065	0.04921049485906564	0.19222176613095404
0.04314257087287088	0.032254392729914094	0.1975260772387093
0.0217480992514882	0.15100084217830906	0.30949786395798945
0.004715035068073295	0.17497794114538537	0.08824966552732103
0.03164967289444218	0.02342110598870864	0.30755254982520713
0.024046679847173927	0.02180549376365151	0.23709064037683017
0.045853716037013055	0.0478490577154183	0.3822384478357849
0.007838745800671874	0.22566663192815048	0.31746861258801207
0.06377084929569174	0.4488120498487927	0.29725180644525456
0.0022985795956857393	0.2068665439776422	0.11748799163424184
0.005245476513231592	0.16470092133771044	0.18036558011000203
0.008074497554075561	0.20159253670126964	0.23419145835776467
0.03329993516826782	0.4583148606298189	0.28278240477066396
0.03164967289444218	0.5472129742978656	0.2776848589793486
0.025932692874403265	0.7167698996923884	0.2720619007819088
0.027759768963281667	0.7432143948365261	0.3459597812633768
0.019803147285907947	1.0	0.4298373759312147

K-mean algorithm is non-adaptive and time consuming and having the accuracy of 65%.

Knearest neighbour

This algorithm is useful if one wants to know the class of given data. First of all the training data must be given with the data and the number of nearest neighbors. On the basis of class of nearest neighbour, this algorithm predicts the possible class of neighbour.

But algorithm gives good result when number of K is more so this algorithm is very time consuming.

Kohonen SOM

BlobSize	BlobAspect	GrayScale
0.03795603229799025	0.0748937051449218	0.5204543791161499
0.8204750397831094	0.024873621055756377	0.5127123845435889
0.09477220486827384	0.012715938827142914	0.5447968461320064
0.07991984440384263	0.060252307458212466	0.5416307692837827
0.15288501208227775	0.10571010643131136	0.41427740926563106
0.023987740908823003	0.0	0.532912008486389
0.030470914127423848	0.010463319810157091	0.32471145316893163
0.056403607001827345	0.0039834255790568915	0.5371541133624109
0.0	0.018829017451228645	0.17545472955420718
0.034537631873637065	0.04921049485906564	0.19222176613095404
0.04314257087287088	0.032254392729914094	0.1975260772387093
0.0217480992514882	0.15100084217830906	0.30949786395798945
0.004715035068073295	0.17497794114538537	0.08824966552732103
0.03164967289444218	0.02342110598870864	0.30755254982520713
0.024046679847173927	0.02180549376365151	0.23709064037683017
0.045853716037013055	0.0478490577154183	0.3822384478357849
0.007838745800671874	0.22566663192815048	0.31746861258801207
0.06377084929569174	0.4488120498487927	0.29725180644525456
0.0022985795956857393	0.2068665439776422	0.11748799163424184
0.005245476513231592	0.16470092133771044	0.18036558011000203
0.008074497554075561	0.20159253670126964	0.23419145835776467
0.03329993516826782	0.4583148606298189	0.28278240477066396
0.03164967289444218	0.5472129742978656	0.2776848589793486
0.025932692874403265	0.7167698996923884	0.2720619007819088
0.027759768963281667	0.7432143948365261	0.3459597812633768
0.019803147285907947	1.0	0.4298373759312147

This algorithm is also used for clustering and quite fast algorithm based of winner take all strategy. it is differentiating the mine and non-mine up to 80% accuracy.

CONCLUSION

As the eight different algorithms have been implemented to compare the results. This classifier is giving result with 80% accuracy. The best result is being given by ART and Genetic algorithm. Fuzzy C mean and gustavson-kessel is also good because of membership values for each class.

This module can differentiate between the PVC tube, wood piece, brass tube, copper cylinder (Non mine data) and the mine data obtained from jrc Israel (<http://apl-database.jrc.it/>).

REFERENCES

- Earl Gose Steve Jost Richard Johnsonbaugh Pattern Recognition and Image Analysis June, 1996 0132364158 Prentice Hall.
- Richard O. Duda ,Peter E. Hart Patter Classification second edition
- Valluru B. Rao C++ Neural Networks and Fuzzy Logic
- Robert catral , franz oppacher Carleton university Ottawa .Rule acquisition with genetic algorithm
- ART Neural Networks for Remote Sensing: Vegetation Classification from Landsat TM and Terrain Data Gail A. Carpenter, Marin N. Gjaja, Sucharita Gopal, and Curtis E. Woodcock.
- <http://polywww.in293.fr/ativities/info/doc/glast/fc.htm>