Scalable Evolutionary Design of Pattern Classifier with Feature Selection Capabilities based on Cellular Automata

Joy Deep Nath Dr. Niloy Ganguly Dr. Pabitra Mitra

Dept. of Computer Science & Engineering
Indian Institute of Technology, Kharagur
18 December 2007

Outline

Outline

Aim of Project

Objective

Design a Scalable Pattern Classifier based on Cellular Automata (*CA*) and Study its Characteristics

Objective 1

Design and implement a Scalable Pattern Classifier based on *CA*

Objective 2

Evaluate the Classifier on datasets of different topologies and experiment to test for Scalability

Objective 3

Experiment on real world datasets and Study the characteristics of the Classifier (Feature Selection)

Aim of Project

Objective

Design a Scalable Pattern Classifier based on Cellular Automata (*CA*) and Study its Characteristics

Objective 1

Design and implement a Scalable Pattern Classifier based on *CA*

Objective 2

Evaluate the Classifier on datasets of different topologies and experiment to test for Scalability

Objective 3

Experiment on real world datasets and Study the characteristics of the Classifier (Feature Selection)

Aim of Project

Objective

Design a Scalable Pattern Classifier based on Cellular Automata (*CA*) and Study its Characteristics

Objective 1

Design and implement a Scalable Pattern Classifier based on *CA*

Objective 2

Evaluate the Classifier on datasets of different topologies and experiment to test for Scalability

Objective 3

Experiment on real world datasets and Study the characteristics of the Classifier (Feature Selection)

Aim of Project

Objective

Design a Scalable Pattern Classifier based on Cellular Automata (*CA*) and Study its Characteristics

Objective 1

Design and implement a Scalable Pattern Classifier based on *CA*

Objective 2

Evaluate the Classifier on datasets of different topologies and experiment to test for Scalability

Objective 3

Experiment on real world datasets and Study the characteristics of the Classifier (Feature Selection)

Outline

Classification Problem and Cellular Automata

Classification Problem

Classification Problem can be viewed as partitioning the feature space into partitions labeled by classes

Machine Learning

ML methods provide technique to determine the boundaries of the partitions in the features space and hence help in learning the classes

MACA Property

MACA, a special class of Cellular Automata partitions the feature space into basins

Classification Problem and Cellular Automata

Classification Problem

Classification Problem can be viewed as partitioning the feature space into partitions labeled by classes

Machine Learning

ML methods provide technique to determine the boundaries of the partitions in the features space and hence help in learning the classes

MACA Property

MACA, a special class of Cellular Automata partitions the feature space into basins

Classification Problem and Cellular Automata

Classification Problem

Classification Problem can be viewed as partitioning the feature space into partitions labeled by classes

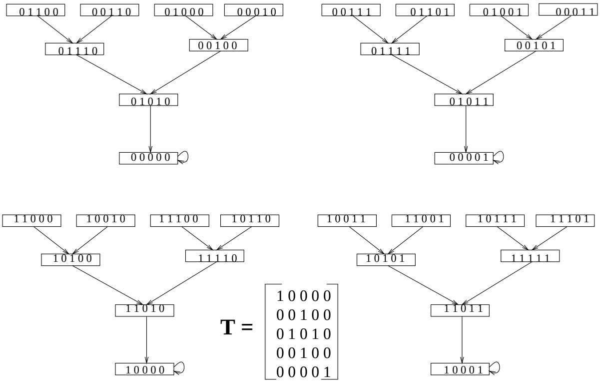
Machine Learning

ML methods provide technique to determine the boundaries of the partitions in the features space and hence help in learning the classes

MACA Property

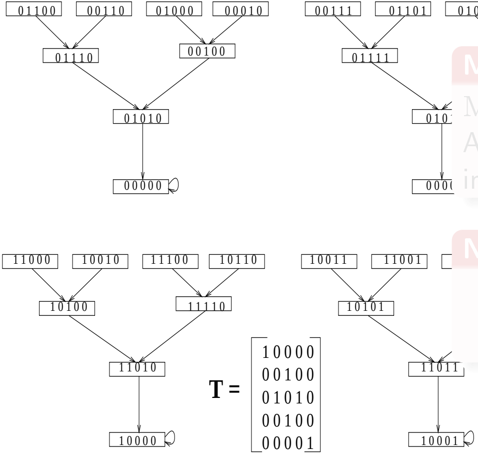
MACA, a special class of Cellular Automata partitions the feature space into basins

MACA



- Given MACA of size n=5, the feature space is of the size 2^n (2^5)=32.
- An MACA is characterized by T Matrix, which captures the basin distribution.

MACA Property



$$\mathbf{T} = \begin{bmatrix} 1&0&0&0&0 \\ 0&0&1&0&0 \\ 0&1&0&1&0 \\ 0&0&1&0&0 \\ 0&0&0&0&1 \end{bmatrix}$$

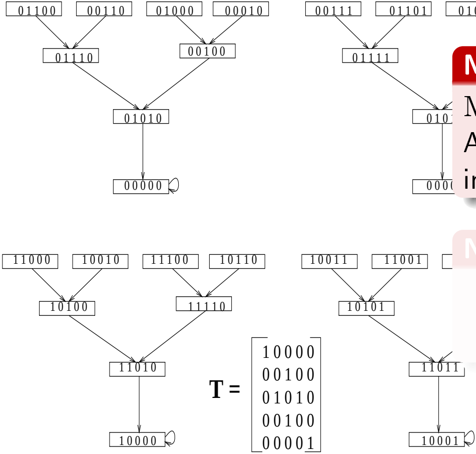
MACA

MACA, a special class of Cellular Automata partitions the feature space into basins

Next State

$$s_{t+1} = T \cdot s_t \;\; i.e, \;\; s_{t+p} = T^p \cdot s_t$$

MACA Property



$$T = \begin{bmatrix} 1&0&0&0&0 \\ 0&0&1&0&0 \\ 0&1&0&1&0 \\ 0&0&1&0&0 \\ 0&0&0&0&1 \end{bmatrix}$$

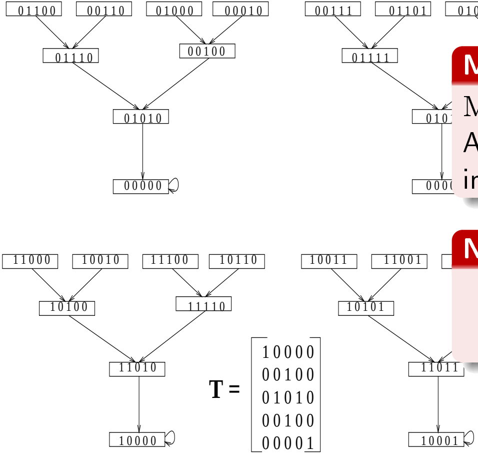
MACA

MACA, a special class of Cellular Automata partitions the feature space into basins

Next State

$$s_{t+1} = T \cdot s_t \;\; i.e, \;\; s_{t+p} = T^p \cdot s_t$$

MACA Property



$$T = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

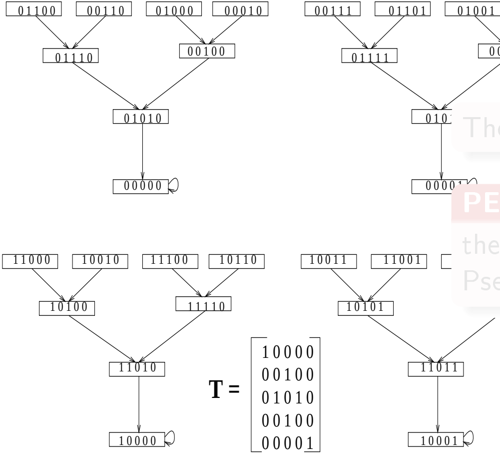
MACA

MACA, a special class of Cellular Automata partitions the feature space into basins

Next State

$$s_{t+1} = T \cdot s_t \;\; i.e, \;\; s_{t+p} = T^p \cdot s_t$$

Attractors



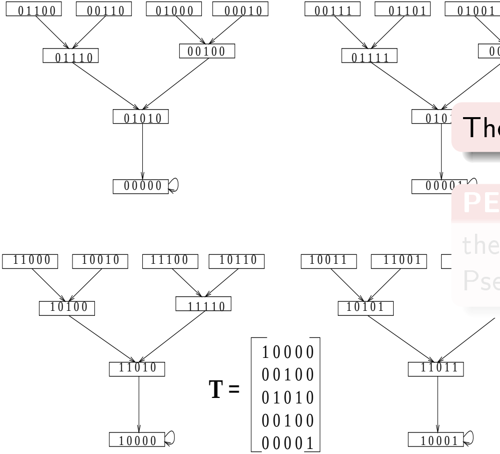
$$T = \begin{bmatrix} 1&0&0&0&0 \\ 0&0&1&0&0 \\ 0&1&0&1&0 \\ 0&0&1&0&0 \\ 0&0&0&0&1 \end{bmatrix}$$

The attractors represent the Class

PEF Bits

the attractors are characterized by Pseudo-Exhaustive Field (*PEF*)bits

Attractors



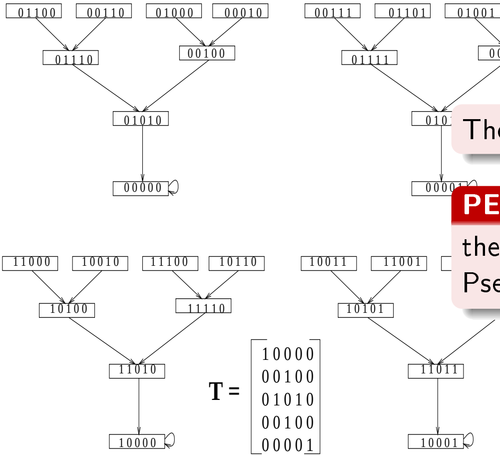
$$
T = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}
$$

The attractors represent the Class

PEF Bits

the attractors are characterized by Pseudo-Exhaustive Field (*PEF*)bits

Attractors



$$T = \begin{bmatrix} 10000 \\ 00100 \\ 01010 \\ 00100 \\ 00001 \end{bmatrix}$$

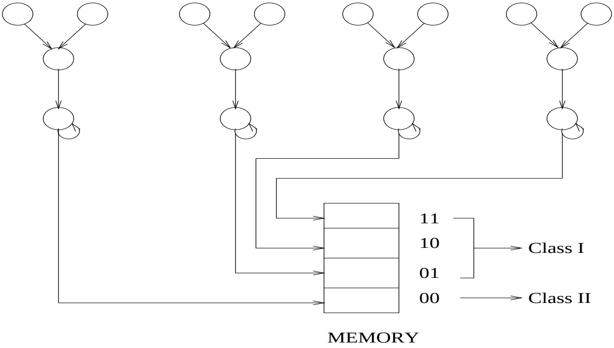
The attractors represent the Class

PEF Bits

the attractors are characterized by Pseudo-Exhaustive Field (*PEF*)bits

Outline

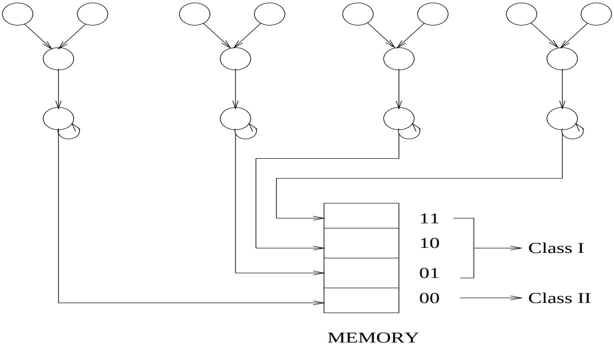
Basic Idea



The features space is divided into basins by the MACA and the basins will be assigned to the classes.

We just need to remember which attactors belong to which class

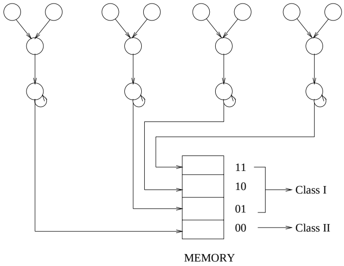
Basic Idea



The features space is divided into basins by the MACA and the basins will be assigned to the classes.

We just need to remember which attactors belong to which class

How it Works



Find an /maca/ that classifies your data and find the attractors corresponding to the clases

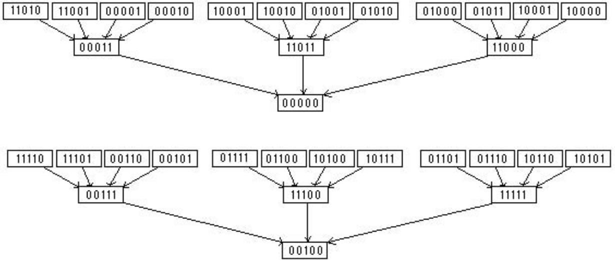
When an incoming pattern comes, Multiply it by the T matrix repeatedly until you reach and attractor

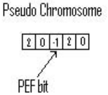
Look up the attractor's class

Learning an MACA

1. We want an MACA that partitions the feature space correctly for us
2. Correctly in the sense that all the patterns belonging to different classes are in different basins
3. We use a GA formulation to search for our appropriate MACA
4. The T Matrix is encoded in a Pseudo-Chromosome and GA is run over it
5. The cost function for GA is calculated over the Training samples based on item 1

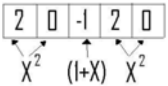
Pseudo-chromosome





$$\begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix}$$

The basin distribution obtained when T Matrix synthesized using Method(I) only



Lot of Matrix Multiplication

In determining which basin a pattern belongs to, you have to repeatedly multiply the pattern with the T Matrix to get the attractor (which characterizes the basin which the pattern belongs to

We would like to have a scheme in which, by just looking at the pattern we can determine the basin/attractor of the class

Lot of Matrix Multiplication

In determining which basin a pattern belongs to, you have to repeatedly multiply the pattern with the T Matrix to get the attractor (which characterizes the basin which the pattern belongs to

We would like to have a scheme in which, by just looking at the pattern we can determine the basin/attractor of the class

Isomorphism in T Matrix

Modifying the algorithm that generates the T Matrix from pseudo-chromosome, we found that a pseudo-chromosome represents an equivalent class of T Matrix which have same Basin Dsitribution

The new scheme enabled us to find out the basin of the pattern by just knowing the PEF bits PEF

In the new scheme we do away with the T matrix altogether, thus saving space

Only the pseudo-chromosome (which contains the position of the PEF bits) is required

Isomorphism in T Matrix

Modifying the algorithm that generates the T Matrix from pseudo-chromosome, we found that a pseudo-chromosome represents an equivalent class of T Matrix which have same Basin Dsitribution

The new scheme enabled us to find out the basin of the pattern by just knowing the PEF bits PEF

In the new scheme we do away with the T matrix altogether, thus saving space

Only the pseudo-chromosome (which contains the position of the PEF bits) is required

Isomorphism in T Matrix

Modifying the algorithm that generates the T Matrix from pseudo-chromosome, we found that a pseudo-chromosome represents an equivalent class of T Matrix which have same Basin Dsitribution

The new scheme enabled us to find out the basin of the pattern by just knowing the PEF bits PEF

In the new scheme we do away with the T matrix altogether, thus saving space

Only the pseudo-chromosome (which contains the position of the PEF bits) is required

Isomorphism in T Matrix

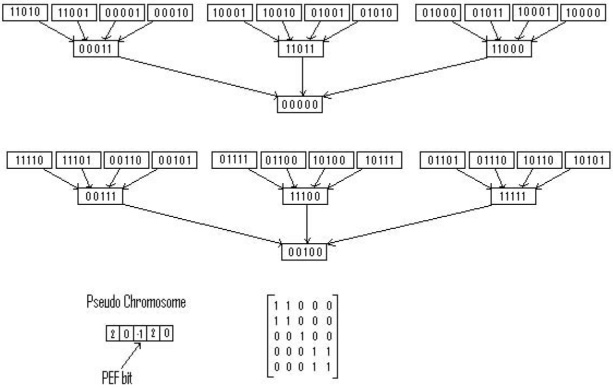
Modifying the algorithm that generates the T Matrix from pseudo-chromosome, we found that a pseudo-chromosome represents an equivalent class of T Matrix which have same Basin Dsitribution

The new scheme enabled us to find out the basin of the pattern by just knowing the PEF bits PEF

In the new scheme we do away with the T matrix altogether, thus saving space

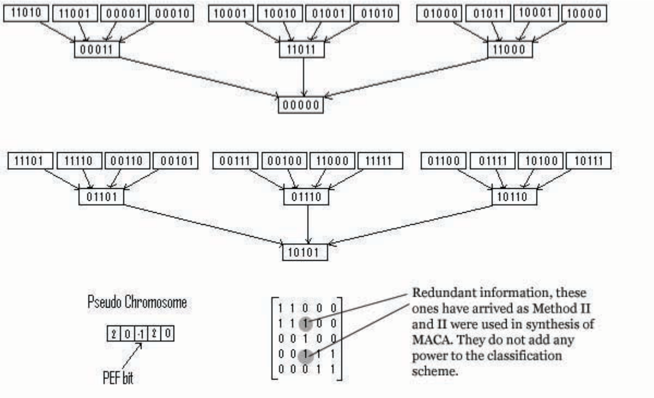
Only the pseudo-chromosome (which contains the position of the PEF bits) is required

Isomorphism



The basin distribution obtained when T Matrix synthesized using Method(I) only

Isomorphism



The Basin distribution obtained when T Matrix is synthesized using Method II and III

Outline

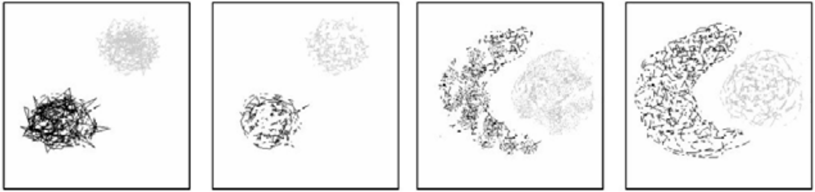
We have a Scalable classifier and we want to see it classify some standard datasets

We artificially create four different topologies of binary classification

1. Linearly Separable Dataset
2. Concave Datasets
3. Spiral Datasets
4. Annular Datasets

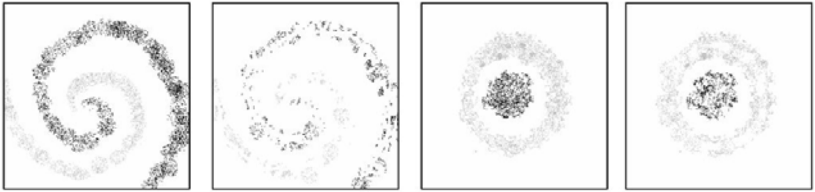
We also compare the performance of the classifier against a Linear Kernel SVM (implementation by Tim Joachims)

Datasets used



training test training test

(a) Linear classification problem (b) Concave Classification Problem



training test training test

(a) Spiral classification problem (b) Annular Classification Problem

Results on the Datsets

Dataset	Scalable *MACA* classifier accuracy on		*SVMLight* Accuracy results
	Training Data	Test Data	
Linear	99.24%	99.61%	99.71%
Concave	92.77%	91.99%	95.44%
Spiral	83.88%	77.45%	82.46%
Annular	73.8%	75.94%	75.95%

Table: Accuracy test results across different Classification problems

Scalability Test

The datasets used previously were small in dimension of feature and small in number of Training Examples

We scaled up both to obtain satisfactory results

Scalability Test

The datasets used previously were small in dimension of feature and small in number of Training Examples

We scaled up both to obtain satisfactory results

Scalability Results

Dataset	Training Data		Test Data		Time taken in seconds
	Accuracy	Examples	Accuracy	Examples	
16	99.24%	4311	99.61%	2313	2
32	98.42%	12000	97.66%	2000	3.0
64	96.22%	50000	96.33%	8000	28
100	100.00%	60000	98.54%	10000	32
100	100.00%	4000	99.69%	2000	3

Table: Scalability Test results

IRS Dataset

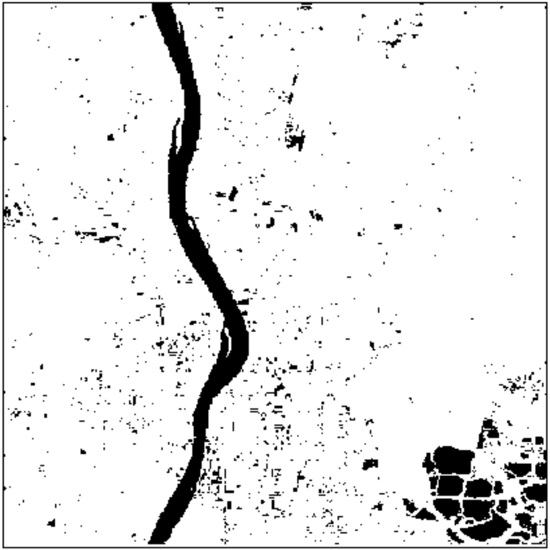
Given

- a 4 band image taken by IRS satellite of Kolkata
- Two classes- (Water body or Man-made construction) or Otherwise
- Take a Set of tagged pixels and learn the classifier and then regenerate the whole image

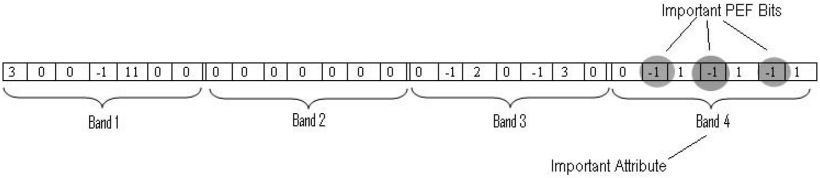
Training Data		Test Data	
accuracy	samples	accuracy	samples
97.76%	120000	97.26%	20000

Table: Accuracy results on IRS Dataset

Reconstructed image



Observation made on the Classifier



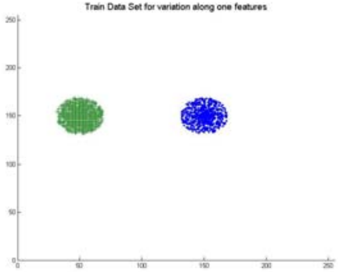
Hints at Feature Selection capabilities

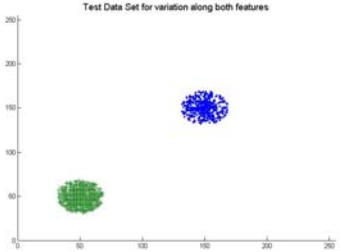
Outline

Two ascertain that there is some sort of feature selection, we perform two more experiments on artificial Datasets. The Dataset had two classes-

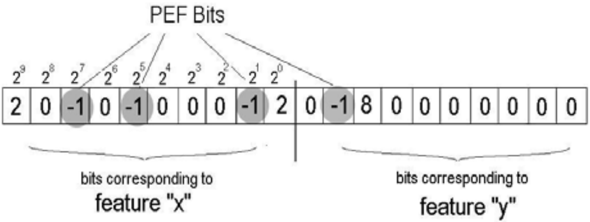
1. Variation along only one feature
2. Variation along both features

Datsets

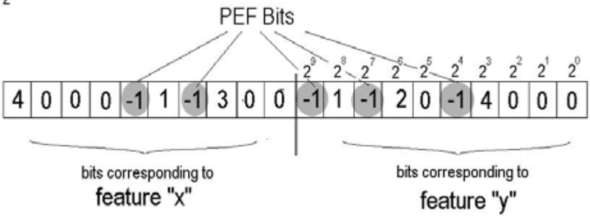

Train Data Set for variation along one features
250
200
150
100
50
0
0
50
100
150
200
250


Test Data Set for variation along both features
250
200
150
100
50
0
0
50
100
150
200
250

Classifiers Learnt Corresponding to the Datsets



2



Feature Selection Observed again

Something more than Feature Selection observed too!!

Feature Selection Observed again

Something more than Feature Selection observed too!!

Gene expression Dataset

A generic approach to cancer classification based on gene expression monitoring by DNA microarrays is described and applied to human acute leukemias as a test case in work done by T. Golub.

- 38 Train samples and 34 Test Samples
- Two classes- ALL and AML
- Each Sample has 7129 features each of whose range can be captured with 19 bits
- Need to classify the Test cases and identify the important genes which the Cancer (Leukemia) Type (ALL or AML) depends.

Golub had medically identified the genes and reported them

Gene expression Dataset

A generic approach to cancer classification based on gene expression monitoring by DNA microarrays is described and applied to human acute leukemias as a test case in work done by T. Golub.

- 38 Train samples and 34 Test Samples
- Two classes- ALL and AML
- Each Sample has 7129 features each of whose range can be captured with 19 bits
- Need to classify the Test cases and identify the important genes which the Cancer (Leukemia) Type (ALL or AML) depends.

Golub had medically identified the genes and reported them

Results

32 out of 34 test cases correctly identified -same accuracy as reported in Golub's work

Features selected were from the important genes reported by Golub!!

Outline

Designed a Scalable Pattern Classifier using cellular Automata.

Extensive experiments done with the proposed classifier

Feature Selection Property observed

Hints at possibility of Rule Generation from the training dataset

Future Work

Dynamics of the classifier to be understood.

Multi-class Classifier based on Decision Tree approach

Investigate Rule generation and inference mechanism based on the classifier as this is of prime interest in data mining

Thank You!!!