Machine Learning in Gene Expression Data Analysis for Cancer Classification using Support Vector Machine

DEVELOPED BY

Memoona Khanam
270-FAS/MSCS/F05

SUPERVISED BY

Prof. Dr. Malik Sikander Hayat Khiyal

Department of Computer Science
Faculty of Basic and Applied Sciences
International Islamic University, Islamabad.
2008

① DNA microarrays

② Gene expression – Research – Methodology

③ Gene expression – Data Processing

④ Genetic algorithms

⑤ BioInformatics

بسم الله الرحمن الرحيم

In The Name of
ALLAH ALMIGHTY
The Most Merciful, The Most Beneficent

Department of Computer Science
International Islamic University, Islamabad

....

FINAL APPROVAL

It is certified that we have read this project report submitted by Miss Memoona Khanam. It is our judgment that this report is sufficient standard to warrant its acceptance by International Islamic University, Islamabad for the MS Degree in Computer Science.

COMMITTEE

1. External Examiner

Dr. A. Sattar
Fmr. Director General
Pakistan Computer Bureau, Islamabad

2. Internal Examiner

Asst Prof., Mr. Shakeel Ahmed
Faculty of Basic & Applied Sciences,
International Islamic University, Islamabad.

3. Supervisor

Hayat

Prof. Dr. Sikander Hayat Khiyal
Chairperson,
Department of Computer Science,
Fatima Jinnah University, Rawalpindi.

A dissertation submitted to the
Department of Computer Science,
Faculty of Basic and Applied Sciences,
International Islamic University, Islamabad, Pakistan
as a partial fulfillment of the requirements for the award of the degree of

MS IN COMPUTER SCIENCES

To

The Holiest Man Ever Born,

PROPHET MUHAMMAD (صلى اللهعيه وسلم)

&

To

MY HOMELAND

Who mean the most to me & recognized me as an individual and independent entity in the world

&

To

MY PARENTS

وقل رب ارحمهما كما ربيانى صغيرا

"And Say: My Lord! Have mercy on them (Parents) both as they did care for me when I was little". Al- Qur'ān 17:23

&

To

My FAMILY

I am most indebted to my parents and family, whose affection has always been the source of encouragement for me, and whose prayers have always been a key to my success.

&

To

THOSE HOLY SEEKERS

Who give away their lives to make the stream of life flow
Smoothly and with Justice.

&

To

MY HONORABLE TEACHERS

Who have been a beacon of knowledge and a constant source of inspiration for my whole life span.

DECLARATION

I hereby declare and affirm that "Machine Learning in Gene Expression Data Analysis for Cancer Classification using Support Vector Machine" software neither as a whole nor as a part thereof has been copied out from any source. It is further declared that I have developed this software and accompanied report entirely on the basis of my personal efforts, made under the sincere guidance of my overseer. If any part of this project is proven to be copied out or found to be a reproduction of some other, I shall stand by the consequences.

No portion of the work presented in this report has been submitted in support of an application for other degree or qualification of this or any other University or Institute of learning.

Memoona Khanam
270-FAS/MSCS/F05

ACKNOWLEDGMENT

I bestow all praises, acclamation and appreciation to الله سبحانه تعالى, The Most Merciful and Compassionate, The Most Gracious and Beneficent, Whose bounteous blessings enabled me to pursue and perceive higher ideals of life, Who bestowed me good health, courage, and knowledge to carry out and complete my work. Special thanks to my Holy Prophet محمد (صلى الله عليه وسلم) who enabled me to recognize my Lord & Creator and brought me the real source of knowledge from الله سبحانه تعالى; the Qur'ān, and who is the role model for me in every aspect of life.

I consider it a proud privileged to express my profound & deepest gratitude and deep sense obligation to my reverend father **Dr. Malik Nazar Hayat Khan** & mother **Mrs. Asifa Nazar Hayat Khan** who kept my moral high by their appreciation. Their motivation and financial support during my entire academic career led me to this success. Without their precious guidance and support I could never ever have been able to accomplish my goal. I owe all my achievements to them, who mean most to me because their prayers are more precious before any treasure on the earth. It will not be out of place to express my profound admiration and gratitude for my supervisor, mentor & guardian **Dr. Malik Sikander Hayat Khiyal** for his dedication, inspiring attitude, untiring help and kind behavior throughout the research efforts and presentation of this manuscript. He did a lot of efforts throughout my graduation & post graduation career.

I am also thankful to my loving grand parents, sisters, brothers, brother in law, relatives, friends, and class fellows who mean the most to me whose prayers have always been a source of determination for me, and to anyone who taught me as a single character since childhood to this stage. May الله سبحانه تعالى blessed all of them with the best of health, admiration at the peak of glory, prosperity, keeps them happy always, fulfill their dreams in their lives & everlasting their sweet & abundant memories in hearts of mankind perpetually (امين).

Memoona Khanam

270-FAS/MSCS/F05

PROJECT IN BRIEF

Project Title:	Machine Learning in Gene Expression Data Analysis for Cancer Classification using Support Vector Machine
Organization	International Islamic University, Islamabad
Under Taken By:	Memoona Khanam 270-FAS/MSCS/F05
Supervised By:	Prof. Dr. Malik Sikander Hayat Khiyal
Starting Date:	February, 2007
End Date:	August, 2008
Tools used:	Matlab 7 & MS Access
System Used	Pentium IV

ABSTRACT

Classification of cancer associated with Philadelphia chromosome is based on the detection of Bcr-Abl gene or ph-chromosome. According to the reciprocal translocation t(9:22)(q34:11) the junction present in the transcript may vary. Recognition of transcript (p120, p210 or p230) does not divulge the type of junction but this information is very important for the classification purpose.

To work out the cancer classification problem by analyzing the gene expression data, in current research we attempt to explore the machine learning classifier method using three benchmark datasets. The classifier is based on the theory of "Support Vector Machine". To systematically evaluate the performance of the machine learning classifier three datasets are CML cancer dataset, AML cancer dataset & ALL cancer dataset. SVM executes the linear kernel function.

Experimental results show that SVM solves the problem of analyzing the cancer related gene expression data and dynamically incorporates with primary database and its performance is directly proportional to the number of trained input samples.

TABLE OF CONTENTS

1. INTRODUCTION 1

2. LITERATURE SURVEY & PROBLEM DOMAIN 3

2.1 LITERATURE SURVEY 3
2.1.1 Machine Learning in DNA Microarray Analysis for Cancer Classification 3
2.1.2 Characteristics of Support Vector Machines in Gene Expression Analysis 4
2.1.3 Using Support Vector Machines for analysis of gene expression data from DNA microarrays 5
2.1.4 Classification of Cancer Tissue Types by Support Vector Machines Using Microarray Gene Expression Data 6
2.1.5 Biological Applications of Support Vector Machines 6
2.1.6 Support Vector Machine Classification of Microarray Gene Expression Data 7
2.1.7 The BCR Gene and Philadelphia Chromosome-positive Leukemogenesis 8
2.1.8 Philadelphia Chromosome–Positive Leukemias: From Basic Mechanisms to Molecular Therapeutics 9
2.2 PROBLEM DOMAIN 10

3. RESEARCH METHODOLOGY 12

3.1 SYSTEM ARCHITECTURE 12
3.1.2 Genetic Expression 14
3.1.3 Feature Selection 14
3.1.3.1 Selection of 22q11 15
3.1.4 Analyzing Expression of 22q11 15
3.1.5 Selection of Bcr-Abl Gene 15
3.1.6 Cancer Classification 15
3.1.6.1 Analyzing Bcr-Abl Gene Sequence 16
3.1.6.2 Classification of Training Sequences 16
3.2 TRAINING ALGORITHM 17
3.2.1 Exemplary Explanation 18
3.2.1.1 Sample Training Input 18
3.2.1.2 Algorithms Execution 18
3.3 APPLICATION PROCEDURE 19
3.3.1 Exemplary Explanation 20
3.3.1.1 Sample Testing Input 20
3.3.1.2 Application Procedure Execution 20

4. IMPLEMENTATION 22

4.1 TECHNOLOGY 22
4.1.1 MATLAB 22
4.1.2 MS Access 22
4.2 SYSTEM'S IMPLEMENTATION 22
4.2.1 Kernel Function for Machine 23
4.2.1.1 Implemented Functions 23
4.2.2 Training of Support Vectors 24
4.2.2.1 Implemented Functions 25
4.2.3 Classification of Support Vectors 28
4.2.3.1 Implemented Functions 29

5. TESTING & RESULTS 31

5.1 TESTING OF SUPPORT VECTOR MACHINE 31
5.2 RESULTS 31
5.2.1 Analysis of Results 32

5.2.1.1 Kernel Function Results 32
5.2.1.2 Training Results 34
5.2.1.3 Modification Results 39
5.2.1.4 Testing Results 43

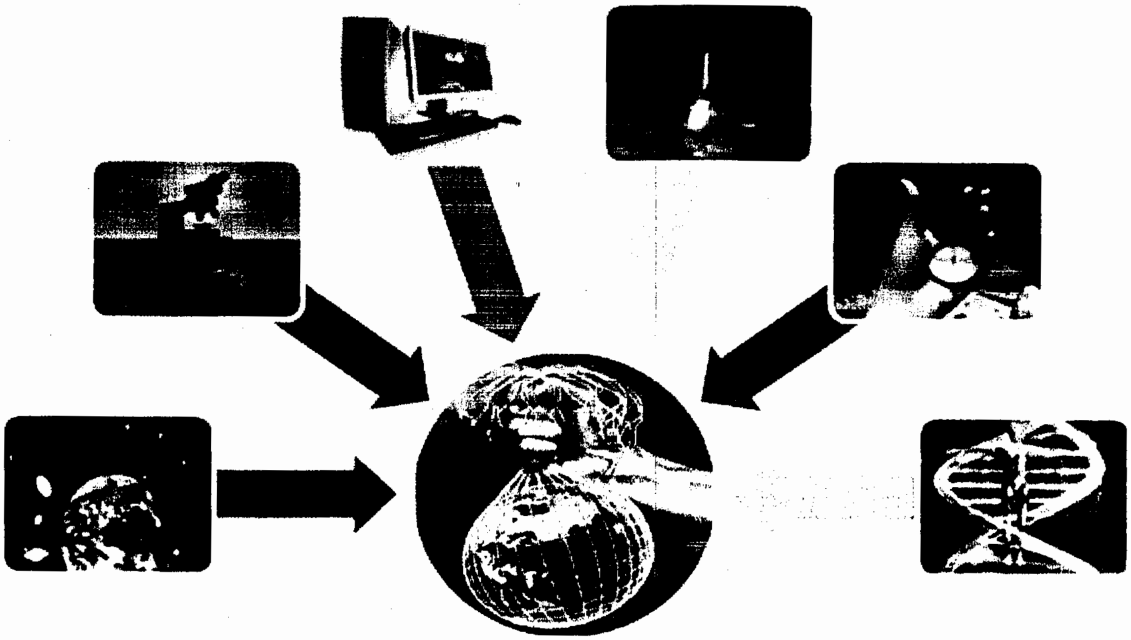
6. CONCLUSION & FUTURE ENHANCEMENT 48

6.1 CONCLUSION 48
6.2 FUTURE ENHANCEMENT 48

REFERNCES 50

CHAPTER 1

INTRODUCTION



1. Introduction

Machine Learning encompasses the automatic computing procedures based on binary or logical operations. The focus of research in machine learning is to extract information from data automatically, by computational and statistical methods.

Support Vector Machines (SVMs) are set of concerned supervised learning methods used for classification. SVMs belong to family of generalized linear classifiers. SVMs simultaneously reduce the empirical classification error and maximize the geometric margin. SVMs are also known as maximum margin classifiers. Support Vector Machines have been employed to classify gene expression data and have shown greater performance than other learning methods, especially in the case where the number of features is larger than the number of samples. (Komura et al. 2002)

One of the major tasks in bioinformatics is the classification and prediction of biological data. With the rapid increase in size of the biological databanks, it is essential to use computer programs to automate the classification process. At present, the computer programs that give the best prediction performance are support vector machines (SVMs). This is because SVMs are designed to maximize the margin to separate two classes so that the trained model generalizes well on unseen data. Most other computer programs implement a classifier through the minimization of error occurred in training, which leads to poorer generalization. (Yang 2004)

Computational Genomics is the study of deciphering biology from genome sequences using computational analysis, including both DNA and RNA. Computational genomics focuses on understanding the human genome, and more generally the principles of how DNA controls the biology of any species at the molecular level. With the current abundance of massive biological datasets, computational studies have become one of the most important means to biological discovery. (2007)

Blood is a highly specialized circulating tissue consisting of several types of cell suspended in a fluid medium known as plasma.

Chromosome is genetic identity and a large macromolecule. DNA normally packaged in a cell in chromosome and gene is the unit of heredity in living organisms. For biological development of a cellular form of life or a virus Deoxyribonucleic acid (DNA) contains the genetic instructions.

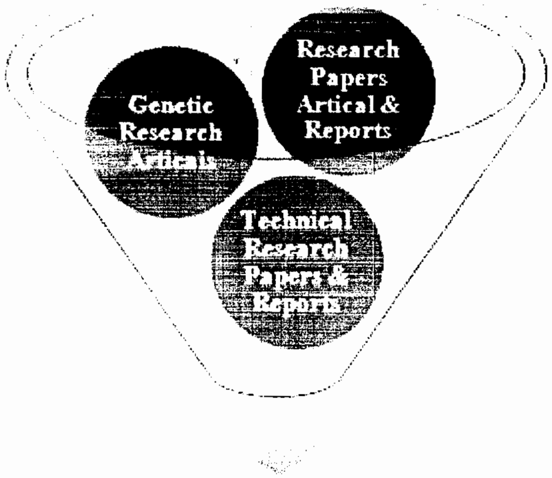
Chronic Myelogenous Leukemia (CML) is a form of leukemia characterized by the increased and unregulated growth of predominantly myeloid cells in the bone marrow and the accumulation of these cells in the blood. CML is a clonal bone marrow stem cell disorder in which proliferation of mature granulocytes (neutrophils, eosinophils, and basophils) and their precursors is the main finding. It is a type of myeloproliferative disease associated with a characteristic chromosomal translocation called the Philadelphia chromosome. (2007)

Philadelphia Chromosome is genetic abnormality associated with a particular form of cancer i.e. CML. The Philadelphia chromosome creates a new leukaemia gene. These abnormal chromosomes are only found in the leukaemia cells. (2007)

To work out the cancer classification problem by analyzing the gene expression data, in current study we endeavor to investigate the machine learning classifier method based on the theory of "Support Vector Machine". We systematically evaluate the performance of the machine by using three datasets includes CML cancer dataset, AML cancer dataset & ALL cancer dataset. Machine executes the linear kernel function that solves the problem of analyzing the cancer related gene expression data.

CHAPTER 2

LITERATURE SURVEY & PROBLEM DOMAIN


Genetic Research Articals
Research Papers Artical & Reports
Technical Research Papers & Reports

Problem Domain

2. Literature Survey & Problem Domain

To produce a healthy research it is essential to comprehensively study the literature that encompasses the research area. Literature survey helps the researcher in analyzing and understanding the available data and shed light on the aims and objectives of the study. Problem domain was clarified after consideration of vast literature.

2.1 Literature Survey

Several research papers were to define the problem domain and achieve the appropriate results of the study. The information retrieved from the selected research papers and articles are given subsequently.

2.1.1 Machine Learning in DNA Microarray Analysis for Cancer Classification

In this paper, Cho et al. (2003) attempt to explore many features and classifiers using three benchmark datasets to systematically evaluate the performances of the feature selection methods and machine learning classifiers. Three benchmark datasets are Leukemia cancer dataset, Colon cancer dataset and Lymphoma cancer data set.

Cho et al. (2003) attempt to explores many features and classifiers that precisely classify cancer using three recently published benchmark dataset. Cho et al. (2003) adopted seven feature selection methods and four classifiers, which are commonly used in the field of data mining and pattern recognition.

Cho et al. (2003) have done cancer classification using procedure shown in figure 5.1

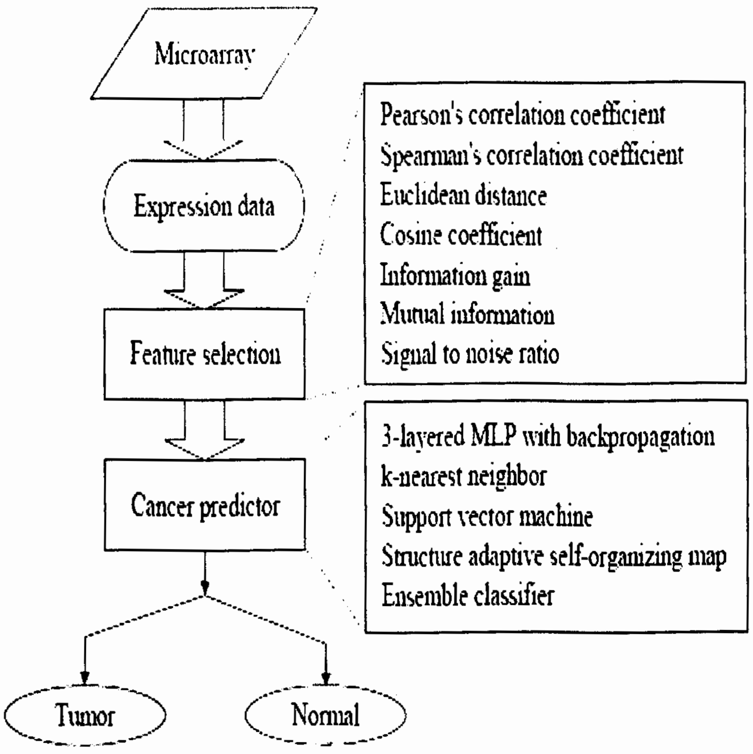


Figure 5.1: Cancer Classification System by Cho et al. (2003)

Results: According to the experiments of Cho et al. (2003) information gain and Pearson's correlation coefficient are the top feature selection methods, and MLP and KNN are the best classifiers.

2.1.2 Characteristics of Support Vector Machines in Gene Expression Analysis

Thus, in this work, Komura et al. (2002) studied how the number of features and regulatable parameters on SVMs affects performance of classification on DNA microarray. Komura et al. (2002) use the DNA microarray data set that consists of two classes; 16 cancerous livers and 8 noncancerous livers obtained from Affymetrix GeneChip. Before analysis, Komura et al. (2002) normalize and filter the raw data. Komura et al. (2002) perform two types of tests. In the first test, they select feature genes at random. In the second test, the feature genes are selected with calculating S/N ratio.

Results: indicate that the linear classification outperforms the nonlinear classifications on the data set. Overfitting causes the misclassification.

2.1.3 Using Support Vector Machines for analysis of gene expression data from DNA microarrays

Fujarewicz et al. (2003) compare the results of different researches. Fujarewicz et al. (2003) presents the comparison of recursive feature replacement (RFR), recursive feature elimination (RFE), Neighborhood analysis (NA) and pure sebestyen methods to the tumor/normal colon and thyroid data sets.

The technique of producing DNA microarrays is improving continuously. In general, there are two different types of DNA microarrays: spotted microarrays and oligonucleotide microarrays. There are several important differences between these two types of microarrays. One of them is the technology of the production. While spotted microarrays are obtained by using special spotting robots, Oligonucleotides microarrays are synthetized, often using photolithographic technology – the same as used during production of computer chips (Fujarewicz et al. 2003).

Fujarewicz et al. (2003) compare the following method of gene selection in their own research

- Recursive Feature Replacement (RFR)
- Recursive Feature Elimination (RFE)
- Neighborhood Analysis (NA)
- Sebestyen Criterion

Results: Fujarewicz et al. (2003) compared RFR, RFE, NA and pure Sebestyen methods to the tumor/normal colon and thyroid data sets. The comparison of obtained results shows that RFR method finds the smallest gene subset that gives no misclassifications in leave-one out cross-validation.

2.1.4 Classification of Cancer Tissue Types by Support Vector Machines Using Microarray Gene Expression Data

To classify cancer and normal tissues Cai et al. (2000) describe a method that based on the gene expression patterns acquired from DNA microarray experiments.

Two supervised machine learning techniques are compared by Cai et al. (2000); one is support vector machines (SVMs) and second is decision tree algorithm (C4.5). ~99% of the tissue samples was correctly classified by SVMs and ~81% of tissue samples able to classify by C4.5

Results: Performance of SVMs significantly better than C4.5.

2.1.5 Biological Applications of Support Vector Machines

Yang (2004) discusses the principles of SVMs and the applications of SVMs to the analysis of biological data, mainly protein and DNA sequences (Yang 2004).

Classification analysis aims to find a mapping function from the features to the class label. There have been many computational algorithms available for the classification analysis of biological data. For instance, decision trees, discriminant analysis, neural networks and support vector machines (SVMs). The essence in classification is to minimize the probability of error in using the trained classifier. This is referred to as the structural risk and it has been shown that SVMs are able to minimize the structural risk through finding a unique hyper-plane with maximum margin to separate data from two classes. Because of this, SVM classifiers provide the best generalization ability on unseen data compared with the other classifiers (Yang 2004).

Many applications of SVMs to biological data analysis are discussed by Yang (2004). Yang (2004) briefly introduces the support vector machines that followed by a discussion of the most important step in using SVMs for analyzing protein or DNA sequences, efficient coding of biological information contained in sequences. Then the applications of SVMs to two classification problems in biology, i.e. modeling whole sequences and subsequences, are discussed (Yang 2004).

Result: According to Yang (2004) there are in general three stages in using SVMs to analyze protein and DNA sequences. In the first stage, the composition and distributed encoding methods are widely used. In the second stage, HMMs are used to generate profiles for families of sequences. From this, profile features are generated. As the profile method uses only positive data leading to weakened prediction accuracy, homology alignment has been used in the third stage. As the homology alignment method for SVMs still has some difficulties in modeling, more advanced methods, particularly advanced kernel functions, are sought for further improvement of the prediction performance (Yang 2004).

2.1.6 Support Vector Machine Classification of Microarray Gene Expression Data

Michael et. al. (1999) introduces a new method of functionally classifying genes using gene expression data from DNA Microarray hybridization experiments. The method is based on the theory of support vector machines (SVMs). Michael et. al. (1999) describe SVMs that use different similarity metrics including a simple dot product of gene expression vectors, polynomial versions of the dot product, and a radial basis function. Compared to the other SVM similarity metrics, the radial basis function SVM appears to provide superior performance in identifying sets of genes with a common function using expression data. In addition, SVM performance is compared to four standard machine learning algorithms. SVMs have many features that make them attractive for gene expression analysis, including their flexibility in choosing a similarity function, sparseness of solution when dealing with large data sets, the ability to handle large feature spaces, and the ability to identify outliers (Michael et al. 1999).

In addition to SVM classification, Michael et. al. (1999) uses following four competing machine learning techniques to analyze the data:

- Fisher's linear discriminant
- Parzen windows
- Two Decision Tree i.e. CART & C 4.5

The SVM method radically gives better results.

2.1.7 The BCR Gene and Philadelphia Chromosome-positive Leukemogenesis

Recent investigations have rapidly added crucial new insights into the complex functions of the normal BCR gene and of the BCR-ABL chimera and are yielding potential therapeutic breakthroughs in the treatment of Philadelphia (Ph) chromosome-positive leukemias.

Laurent et. al. (2001) comprehensively gives the molecular features of BCR gene in table 5.1, ABL gene in table 5.2

Table 5.1 BCR: Molecular Features

Feature	Comments
Location	22q11 (On short arm of chromosome 22)
Size of gene	130-kb
Number of exons	23 exons (also contains alternative exon 1 and exon 2)
Size of transcripts	4.5kb and 6.7kb
Size of proteins	M_r 130,000 and M_r 160,000[a]
BCR expression	• Ubiquitously expressed • Highest levels in brain and haematopoietic cells

BCR-related genes	Pseudogenes BCR2, BCR3, and BCR4 found on chromosome 22q11; ABR (active BCRrelated) gene on 17p13

Table 5.2 ABL: Molecular Features & Function

Feature	Comments
Location	Chromosome 9q34 (On short arm of chromosome 9)
Size of gene	>230 kb
Number of exons	11 exons (There are two alternative exons, 1a and 1b)
Size of transcripts	6.0-kb and 7.0-kb
Size of proteins	M_r 145,000

The hallmark of CML is the Ph chromosome, which is a shortened chromosome 22 resulting from a translocation, t(9;22)(q34;q11), between chromosomes 9 and 22.

2.1.8 Philadelphia Chromosome–Positive Leukemias: From Basic Mechanisms to Molecular Therapeutics

The Philadelphia chromosome translocation (t(9;22)) results in the molecular juxtaposition of two genes, Bcr and Abl, to form an aberrant Bcr-Abl gene on chromosome 22.

Kurzrock et al. (2003) present a comprehensive overview of the molecular genetics of the philadelphia chromosome translocation. Kurzrock et al. (2003) presents the schematic diagram of normal Bcr, Abl proteins and various aberrant Bcr-Abl counterparts in figure 5.2.

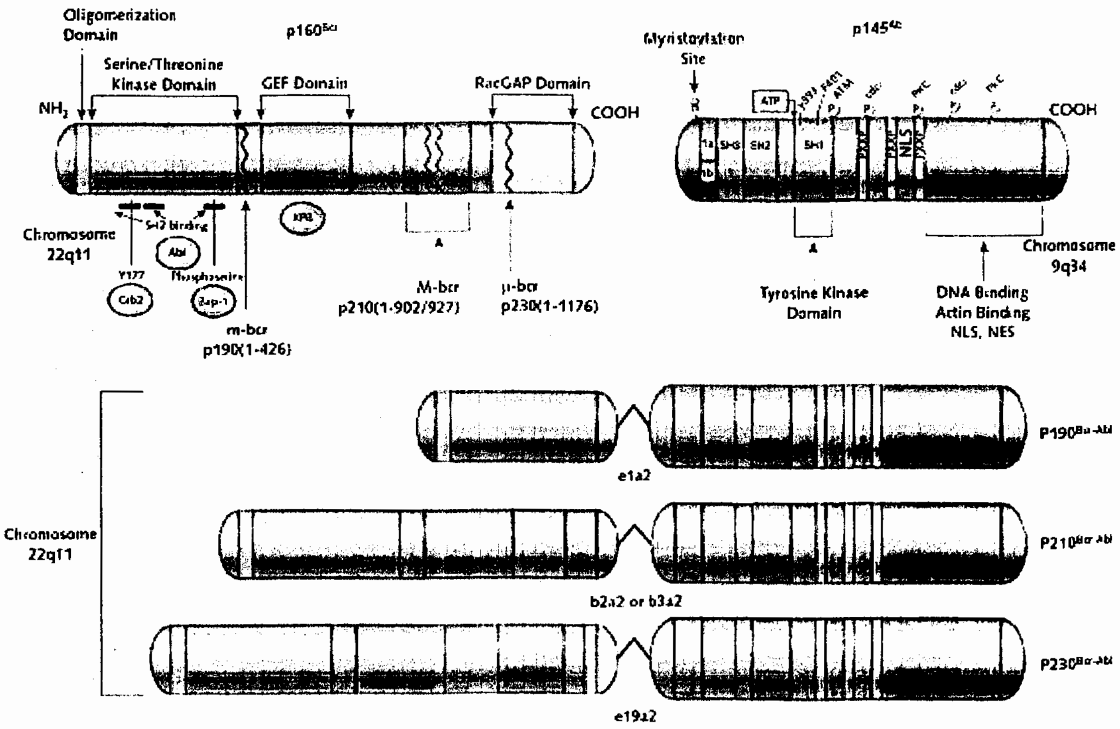


Figure 5.2: The Normal Bcr and Abl Proteins and the Various Aberrant Bcr-Abl Counterparts

2.2 Problem Domain

Recent years have seen dramatic and sustained growth in the amount of genomic data being generated, including in late 1999 the first complete sequence of a human chromosome. The challenge now faced by biological scientists is to make sense of this vast amount of accumulated and accumulating data.

Many biological researchers have been studying many problems of cancer classification using gene expression profile data and attempting to propose the optimal classification technique to work out these problems. Still there has been no comprehensive work to compare the possible feature selection methods and classifiers for gene expression data. The gene expression data usually consist of huge number of genes. The key problem in analyzing and classifying this genomic data is how to integrate the software and primary databases in a flexible and robust way.

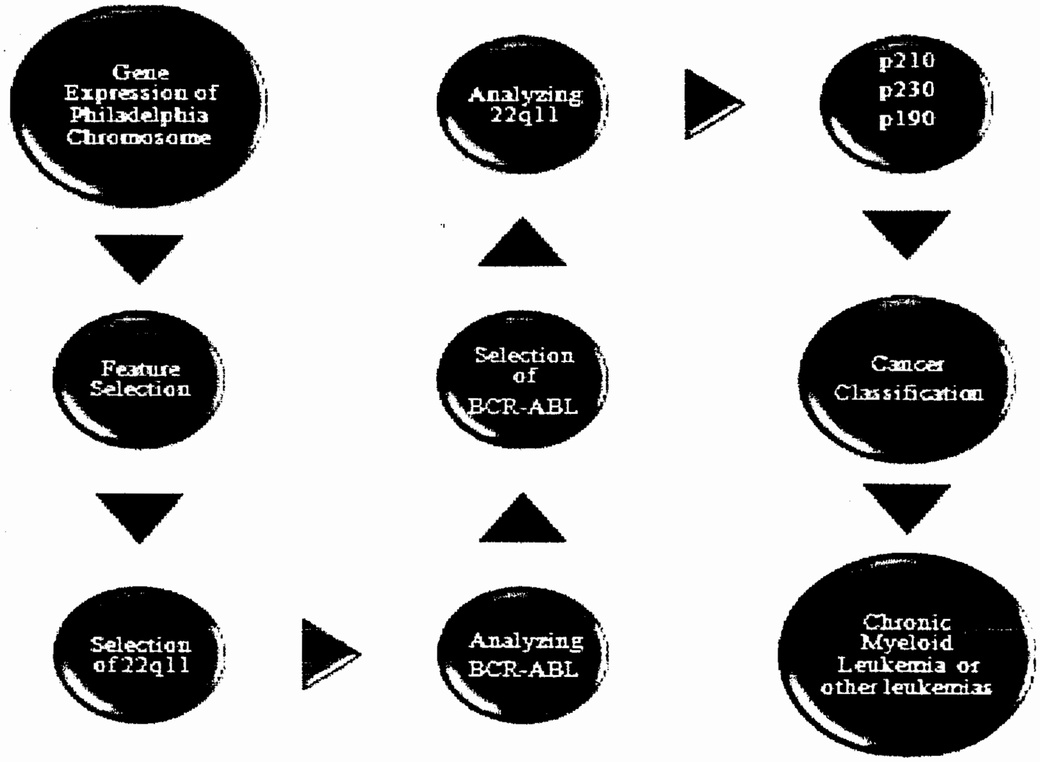
The aim and objectives of the study are:

- To make sense of vast amount of accumulated and non-accumulated biological data
- Use of mathematical tools to extract useful information from accumulated data produced by high-throughput biological techniques such as genome sequencing
- Investigation of genetic sequence of required gene and comprehensively classification of accumulated data by integrating machine learning techniques and basic databases
- To analyze and correctly classify the two classes of leukemogenic cancers using the "Support Vector Machine". These two classes of cancer include CML and non-CML leukemogenesis. For this purpose genetics of these two classes are comprehensively studied.

On the basis of literature survey, the research methodology is presented in next chapter.

CHAPTER 3

RESEARCH METHODOLOGY



3. Research Methodology

According to the objective of study, the current research work is laying under the one of major research types known as "Fundamental Research".

In this study, it is desired to develop an intelligent decision system that investigate and comprehensively classify the genetic sequence of required gene i.e. BCR-ABL in two classes i.e. +ve & -ve. Positive class contains the training samples of CML and negative class is trained with other leukemogenesis i.e. ALL & AML. Finally it is requisite that some testing samples of both classes are correctly classified that satisfy the constraints and conditions of the system.

The initial sequencing of the human genome and recent technological advances now make it possible to develop global views of the cell at a molecular level. The molecular analysis using pattern recognition is especially relevant in cancer genomics.

3.1 System Architecture

"The architecture of system is comprehensive framework that describes its forms and structure, its components and how they fit together". (Pressman 2001) The architecture is not the operational software. Rather, it is representation that enables a software engineer to:

- Analyze the effectiveness of the design in meeting its stated requirement.
- Consider architectural alternative at a stage when making design changes is still relatively easy.
- Reducing the risk associated with the construction of the software.

Keeping in view the performance provided by the tools/techniques empowered with intelligent algorithms and presence of limited literature on Support Vector Machine: A Statistical Learning Theory regarding cancer classification: this research proposes a cancer classification methodology based on Support Vector Machine: A Statistical Learning Theory. It aims to design a technique that offers better accuracy, focuses on offering higher efficiency. Work flow of proposed system is shown in figure 6.1.

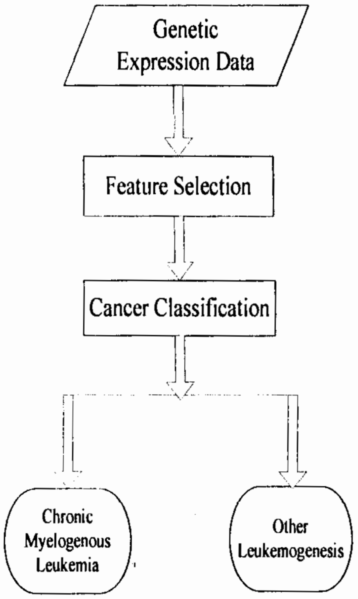


Figure 6.1: Work Flow of Proposed System

There are three main modules and 6 sub modules cover in the current research. Due to multidimensional domain the research required more then prescribed time to implement the all modules. Also they required the complete and comprehensive knowledge of gene transcription, translation, translocation and the functionality they perform during and after above each process the first two modules are hard-code implemented and the third module is soft-code implemented. Figure 6.2 is the architecture diagram representing the flow of modules cover in study.

- Genetic Expression of Philadelphia Chromosome
- Feature Selection
  - Selection of 22q11
  - Analyzing Expression of 22q11
  - Selection of Bcr-Abl Gene
- Cancer Classification
  - Analyzing Bcr-Abl Gene Sequence

- Classification of training sequences
- Testing based on trained classified sample data

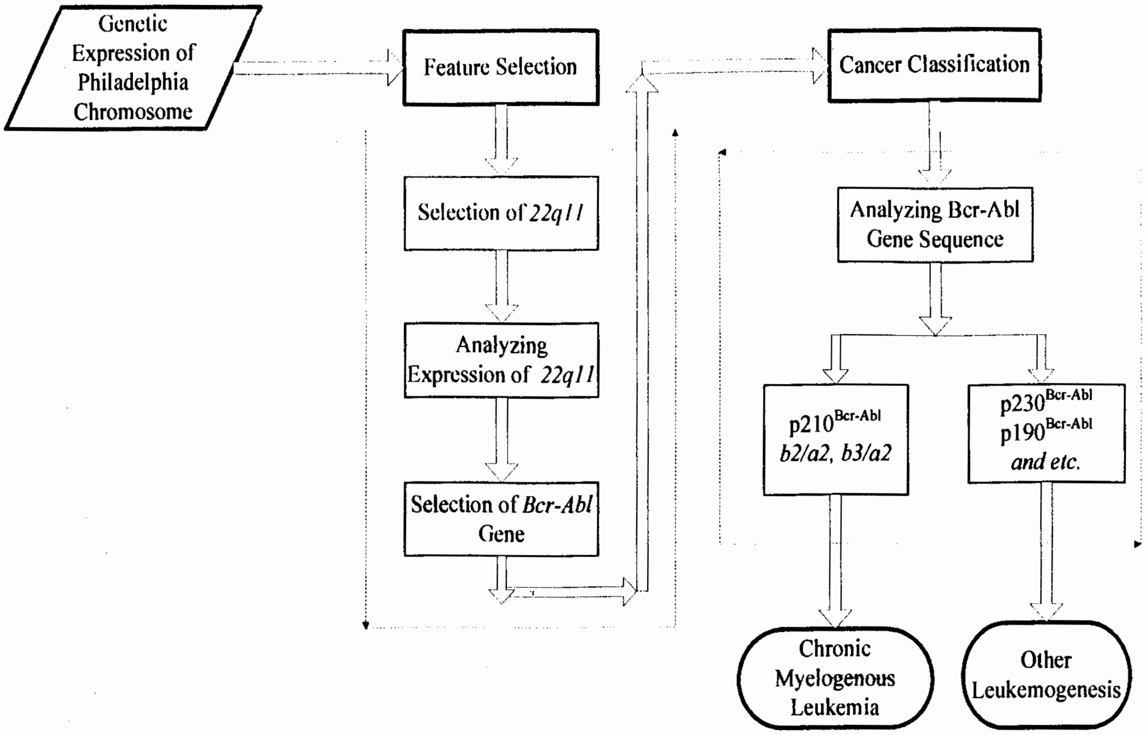

Genetic Expression of Philadelphia Chromosome
Feature Selection
Selection of 22q11
Analyzing Expression of 22q11
Selection of Bcr-Abl Gene
Cancer Classification
Analyzing Bcr-Abl Gene Sequence
p210Bcr-Abl b2/a2, b3/a2
p230Bcr-Abl p190Bcr-Abl and etc.
Chronic Myelogenous Leukemia
Other Leukemogenesis

Figure 6.2: Architecture Diagram of Proposed System

3.1.2 Genetic Expression

In current research gene expression of Philadelphia chromosome is analyse which is the chromosome abnormality that causes chronic myeloid leukemia (CML). Hallmark of chronic myeloid leukemia is Philadelphia chromosome. In said research genetic expression of ph-chromosome is analyzed to achieve our required gene information.

3.1.3 Feature Selection

In situations where there is an abundance of training data it is conceivable that the training of classifiers cannot be performed over the full set of data due to limited computing resources. Thus a question arises how the training of a given classifier can still be done so that the full training set is taken into account.

Feature selection is distinct as "the search for a subset of the original measurement features that provide an optimal trade off between probability error and cost of classification". It involves selecting a subset from the original set of features that captures the relevant properties of the data to enable adequate classification.

3.1.3.1 Selection of 22q11

22q11 is the portion of Philadelphia chromosome where the Bcr-Abl is located and performing the functionality that causes the chronic mylogenous leukaemia. In this sub module the portion of Philadelphia chromosome is selected for analysis.

3.1.4 Analyzing Expression of 22q11

The expression of selected portion of Philadelphia chromosome (22q11) is analyzed to find out the point and position of translocated Bcr-Abl on that chromosome.

3.1.5 Selection of Bcr-Abl Gene

By analysing the expression of 22q11 the translocated Bcr-Abl is selected that gives us the selected gene location point, its expression that containing the prescribed genetic code of required gene.

3.1.6 Cancer Classification

Over the last few years the rapid growth in the field of genomics has made possible the creation of large data sets of genetic information characterizing complex biological systems. Molecular classification approaches based on machine learning algorithms applied to genetic data have been shown to have statistical and clinical relevance for a variety of tumor types: Leukemia, Lymphoma, Brain cancer, Lung cancer and the classification of multiple primary tumors.

In the present research, we use one particular state-of-the-art machine learning classification algorithm, Support Vector Machines (SVMs). SVMs are prevailing classification systems based on regularization techniques with outstanding performance in many practical classifications. The technique developed an intelligent machine that uses small subset of highly discriminant genetic data (Gene Sequence) to build very reliable cancer classifiers.

General process of classification in machine learning is to train classifier to accurately recognize patterns/sequences from the given training samples and to classify test samples with the trained samples.

3.1.6.1 Analyzing Bcr-Abl Gene Sequence

Analysis of Bcr-Abl gene sequence is furnished to check that weather it causes the CML or other leukemogenesis. This analysis is done by the molecular weight of selected gene. The gene causes the CML have $p210^{Bcr\text{-}Abl}$ molecular weight and the gene containing the $p190^{Bcr\text{-}Abl}$ & $p230^{Bcr\text{-}Abl}$ weight causes the other leukemogenesis i.e. ALL & AML

3.1.6.2 Classification of Training Sequences

Genetic sequence of Bcr-Abl gene of many patients are used to train the intelligent system that classify the CML leukemogenesis and non-CML leukemogenesis in two different classes based on the label that we give to each input of both classes. The label of inputs belongs to CML class is selected +1 and the other belongs to non-CML is -1. For classification of two different types of leukemogenic classes is furnished using the "Support Vector Machine". The training algorithm is presented subsequently.

6.1.3.1 Testing of Cancer Classification System

The testing of cancer classification system is done by taking some sample inputs to test that whether the system correctly classifies the samples or not. Testing is based on trained classified sample data. Testing procedure is presented subsequently.

3.2 Training Algorithm

```
Step 0:
        Initialize weight and bias
Step 1:
        While stopping condition is false do steps 1-6
Step 2:
        For each training pair ($X_i$, $Y_i$) do steps 3-5
Step 3:
        Set activation of input unit
                $X_i = S$                    // s: Input sequence of gene
Step 4:
        Set activation of label to input unit
        where
                $Y_i \varepsilon \{1, -1\}$
Step 5:
        Compute activation of output unit
                $F(X) = \sum_{i=1}^{m} \alpha_i Y_i K(X, X_i) + b$
        $Y_i \varepsilon \{1, -1\}$
        $X_i \varepsilon R^n$ (Support vectors)
        $\alpha_i$      Non zero SV or weight
        b       bias
        K(.)    Kernel function i.e. linear kernel function
Step 6:
        Test for stopping condition
        If
        no input sequence is available in step 2
        else
        continue
```

3.2.1 Exemplary Explanation

Paper execution of training algorithm with sample training input sequence is given subsequently.

3.2.1.1 Sample Training Input

The sample training input sequence is given below. This is sample input sequence. Original training sample may contain thousands of bases.

GTATCATGCCTATTACATAGTAGATTTTCAATAAATGTTAAATGCATGAATGA
TGGGTAATGTTTTTAGCTTGGAGGAGAGAAGCCTAAGGGAAGATGTGTTTGC
TGTCGCTGGGCGCGGTGGCTCACGCCTGTAATCCCAGCACTTTGGGAGGCTG
GGTCAGGAGTTCGAGACCAGTCTGGCCAACATGGTGAAACCCCGTCTCTACT
AGGAGGCGGAGGTTGCAGTGAGCCAAGATTGCACCACTGCACTCCAGTCTGG
GGCCGAGGCGGGTGGATTACGAGGTCAGGAGATCGAGACCATCCTGGCTAAC
TACTAAAAATACAAAAAATTAGCCAGGCGTGGTGGCGGGCACCTGTAGTCCC
CCAGTGATTATTGGTCATTGGAGGTATTAAAGGAGAGGCAAGATAGGAAGAT
TGTTAGGAATGGCTTGGCTAGGATAGTCAGAGGAGTACGTGGAGTAAGGCAG

3.2.1.2 Algorithms Execution

Algorithms generally executes the linear kernel function that generates two decision surfaces for two required classes in a feature space and these decision surfaces are separated by decision boundary. The decision surfaces and feature space are dynamically changed with the number of training input vectors.

Step 0:

Weight and bias are set as 0.5

Step 1:

Until the stopping condition is false do steps 1-6. If there is no input vector available its mean that algorithms meets its stopping condition.

Step 2:

For each training pair (X_i, Y_i) the machine can do steps 3-5, where X_i is the input vector for training and Y_i defined set of labels

Step 3:

At this step the input is given to machine for training. This input vector is in form of text file that contains the genetic information about the different types of cancer associate with ph-chromosome. Sample data is shown previously.

Step 4:

The input vectors are labelled with available labels. The vectors belongs to the CML class are labelled with 1 and other one are labelled with -1.

Step 5:

The machine executes a formula and train the labelled input vectors in into respective class. The machine represents the trained vectors as the dot/point in respective decision surfaces.

Step 6:

Stopping condition is tested to terminate the program and that is no sequence /vector available for training

3.3 Application Procedure

Step 0:

Apply training algorithm

Step 1:

For each input vector X to be classified do steps 1-3

Step 2:

Set activation of input units

Step 3:

Compute response of output units

3.3.1 Exemplary Explanation

Paper execution of application procedure with sample input sequence is given subsequently.

3.3.1.1 Sample Testing Input

The sample testing input sequence is given below. This is sample testing sequence. Original sample may contain thousands of bases.

TGGGTAATGTTTTTAGCTTGGAGGAGAAGCCTAAGGGAAGATGTGTTTGC
GGTCAGGAGTTCGAGACCAGTCTGGCCAACATGGTGAAACCCCGTCTCTACT
GCTGGGGGTAGTGGCGTTGCCTATAATCTCAGCTACTTGGGAGGCTGAGGCA
AGGAGGCGGAGGTTGCAGTGAGCCAAGATTGCACCACTGCACTCCAGTCTGG
GGCCGAGGCGGGTGGATTACGAGGTCAGGAGATCGAGACCATCCTGGCTAAC
TACTAAAAATACAAAAAATTAGCCAGGCGTGGTGGCGGGCACCTGTAGTCCC
CCAGTGATTATTGGTCATTGGAGGTATTAAAGGAGAGGCAAGATAGGAAGAT
TGTTAGGAATGGCTTGGCTAGGATAGTCAGAGGAGTACGTGGAGTAAGGCAG

3.3.1.2 Application Procedure Execution

After training SVM can be used to classify the input vectors.

Step 0:

Training algorithm is applied until all the sample are trained

Step 1:

For each input vector X to be classified do steps 1-3

Step 2:

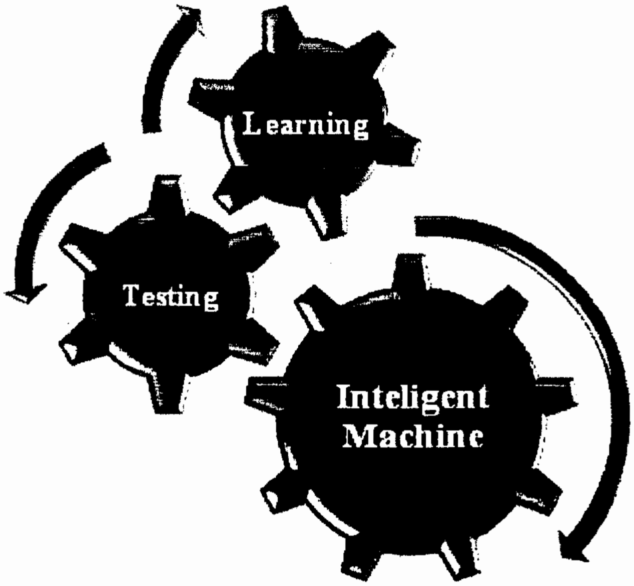
At this step the input is given to machine for classification. This input vector is in form of text file that contains the genetic information about the different types of cancer associate with ph-chromosome. Sample data is shown previously.

Step 3:

For classification purpose contents of testing input vector is compare character by character with all trained vectors. If the test vector is matched with one of the training input then the respective trained vector is blinked in respective decision surface. But in case, the test vector is not match with the trained vector then none of trained vectors blink in both decision surfaces.

CHAPTER 4

IMPLEMENTATION



4. Implementation

Implementation is an important stage and phase of software lifecycle where the thoughts and ideas are presented in corporal structure. Implementation is a depiction of the note-worthy organization of deliverables. A good implementation approach and strategy leads to achievement of goal i.e. successful intelligent machine.

4.1 Technology

For the purpose of implementing the intelligent machine following tools are used in the current research.

- MATLAB 7
- MS-Access
- The application can run on Pentium IV with at least 2.8GHz processor on Windows operating system.

4.1.1 MATLAB

MATLAB; A Matrix Laboratory is a high-performance tool of choice for high-productivity research. MATLAB featured with toolboxes. Toolboxes contain the comprehensive collections of MATLAB functions (M-files) that extend the MATLAB environment to solve particular problems.

4.1.2 MS Access

Effective database management must begin by thinking about how people use the information. For creation and management of databases MS Access is a powerful program.

4.2 System's Implementation

Main modules of the learning machine are as following:

- Kernel function for machine
- Training of support vectors
- Classification of support vectors

4.2.1 Kernel Function for Machine

The SV method describes the general concept of learning machine. It considers a kernel function approximation that has to satisfy the two conditions:

i. The kernel function that defines the SV machine has to satisfy some pre-define conditions.
ii. The hyperplane constructed in the feature space has to be optimal.

In current research we use the linear kernel function.

4.2.1.1 Implemented Functions

Following function is used for implementation of kernel function:

A. linear()

This function defines the SVM and satisfies the equation of line. It constructs a linear hyperplane as decision boundary in a prescribed feature space. Using this function feature space can be increase and decrease as per user & available data requirement.

Feature space having decision boundary is shown in figure 7.1 & extended feature space having decision boundary is presented in figure 7.2. Feature space can be extended either before starting any trainings or during training of sample input. The function not only construct the linear hyperplane but also keeps records of all the points either initial or extended that are being plotted to specify a decision boundary.

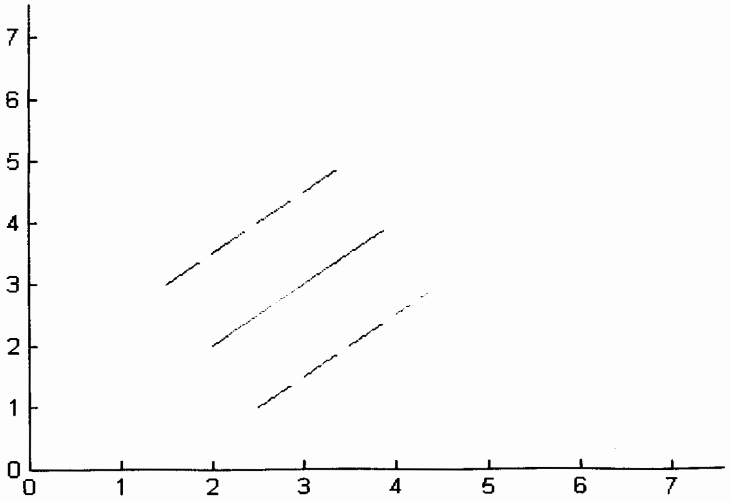


Figure 7.1: Feature Space Having Decision Boundary

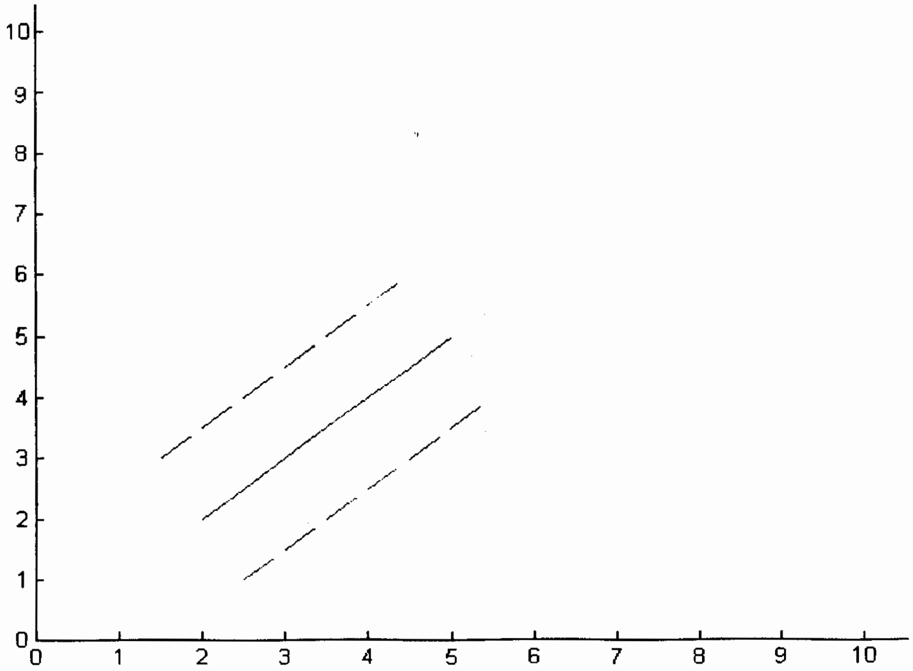


Figure 7.2: Extended Feature Space Having Decision Boundary

4.2.2 Training of Support Vectors

Genetic sequence of Bcr-Abl gene of many patients are used to train the intelligent system that classify the CML leukemogenesis and non-CML leukemogenesis in two different classes based on the label that we give to each input of both classes. The label of inputs belongs to CML class is selected +1 and the other belongs to non-CML is -1. For classification of two different types of leukemogenic classes is furnished using the "Support Vector Machine". The training algorithm is presented in section 6.2.

4.2.2.1 Implemented Functions

Subsequent implemented functions are used to train the SVM:

A. svtrain()

This function train the sample input vector into two classes based on the label given to each input vector of both classes. The input vectors are the text file that contains the sample genetic sequences of different patients. The classes in which input vectors are train to correctly classify the vectors includes CML & non-CML.

The function takes the name and label of file belongs to classes and train it into respected class and insert the file into respected database table. The training is successfully done only when the function detects that text file is present of hard disk. If the input file is not present on the hard drive the firstly it request to put the file on hard drive and then train it. The trainings of input sample are proceeds one by one that is represented on figure 7.3.

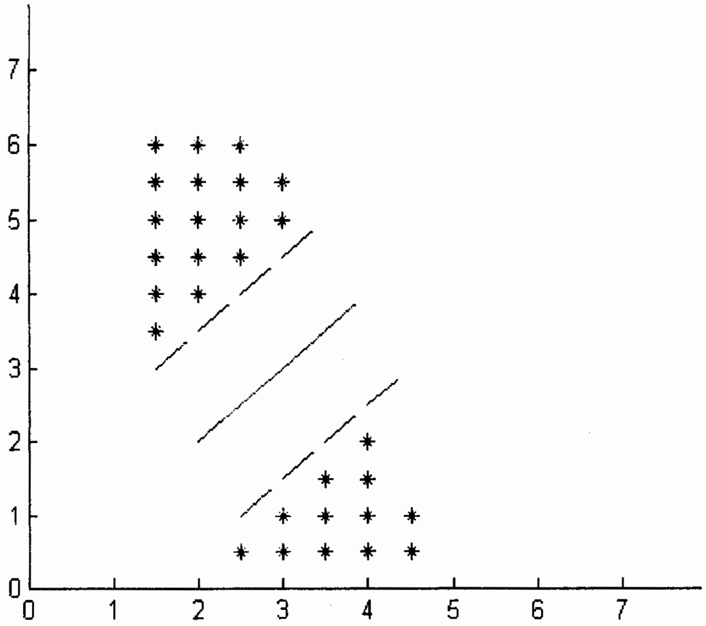


Figure 7.3: Training of Input samples

The function not only trains the vectors but also reduces the risk of error that occurs due to un-availability of file during testing. The function automatically increases the decision boundary by increasing the input vectors.

The function performs following dynamic executions; that are:

i. Feature can be extended during training as per requirement of machine training engineer. Extended feature space of figure 7.3 is shown in figure 7.4

ii. Decision surface of each class can be increased during training.

iii. During training the decision surface can be increased and decreased as per data availability and user requirement.

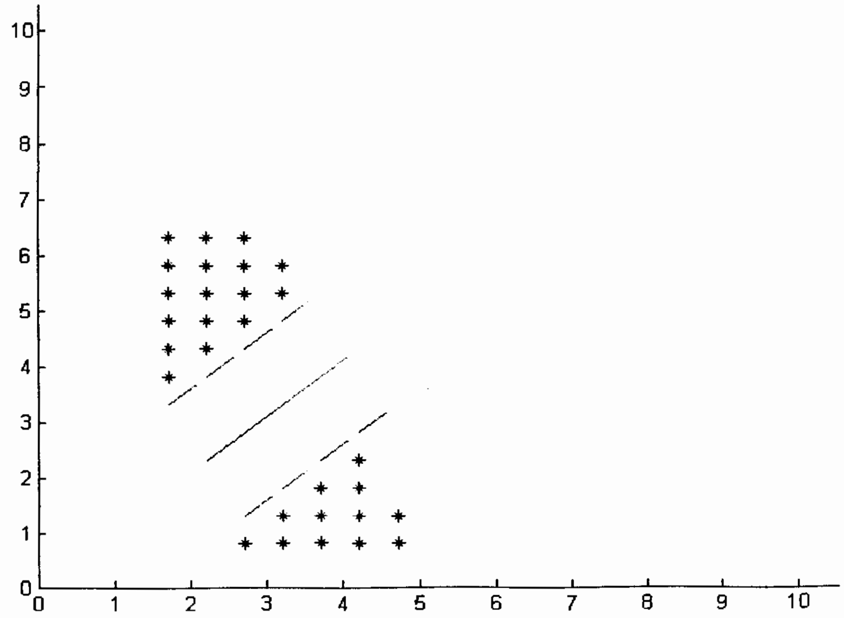


Figure 7.4: Extended Feature Space During Training

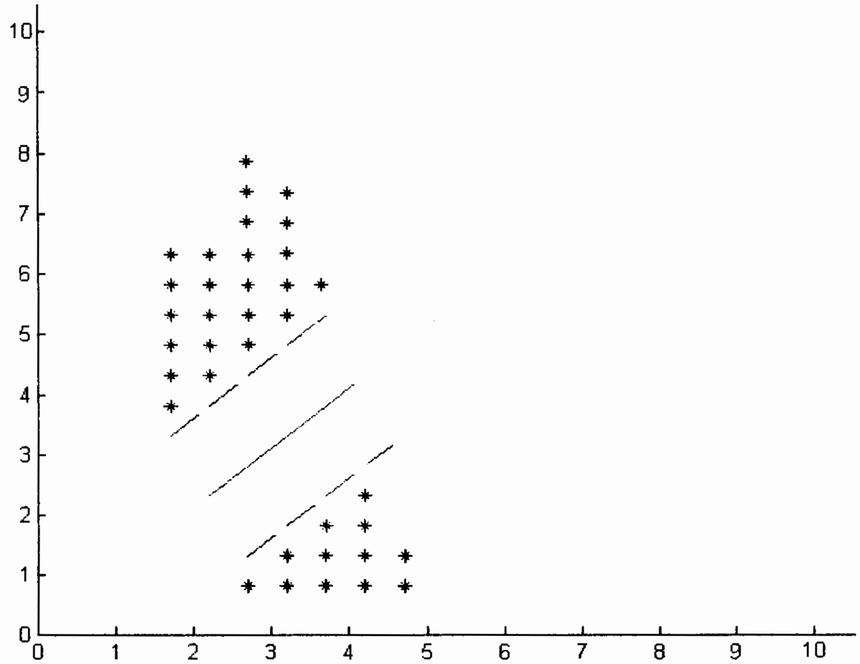


Figure 7.5: Changing the Bounds of Decision Surface During Training

B. svmod()

During training there is a probability of training the wrong file or the trained may contain outdated data. To fix this problem svmod() is introduces in the machine that delete the wrong trained file from the training system.

The modifications of five input samples i.e. shown in figure 7.3 are processed one by one and decreases the decision boundary that is represented in figure 7.6.

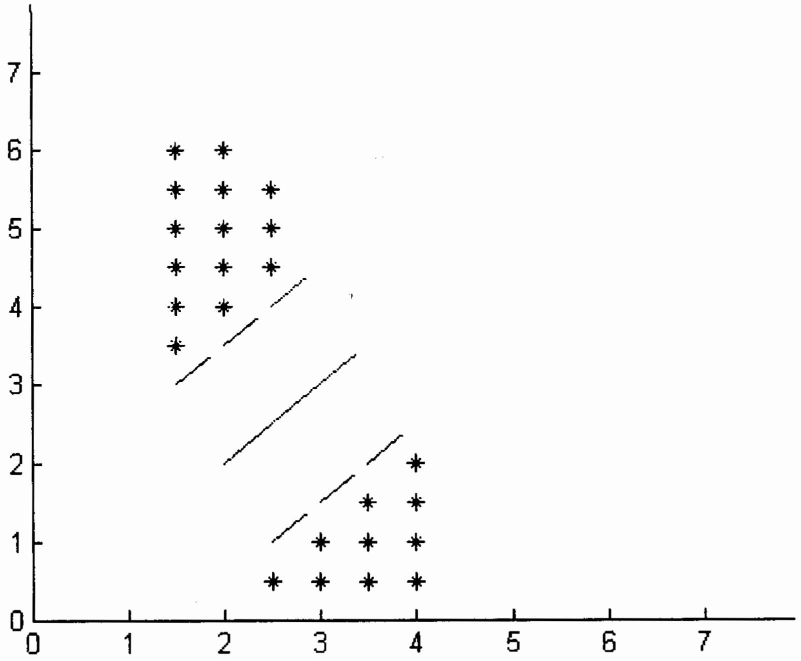


Figure 7.6: Modification of Training Samples

The function reduces the risk of error that occurs due to wrong training of input vector. The function automatically decreases the decision boundary by decreasing the sample vectors.

4.2.3 Classification of Support Vectors

The testing of cancer classification system is done by taking some sample inputs to test that whether the system correctly classifies the samples or not. Testing is based on trained classified sample data. Testing procedure is presented in section 6.3.

4.2.3.1 Implemented Functions

Subsequent implemented functions are used to test the SVM:

A. svtest()

This function takes the sample vectors and classifies them in either to belonging classes. The function takes the name of sample vector from user and proceeds to ensure:

i. Availability of file on hard drive
ii. Availability of file in trained sample

If the function doesn't detect the file with that name then it firstly requested to insert the test sample's text file on the hard drive. If function detect the test file from the hard drive then it compare the contents of that file with all the trained vectors and classify it into respected class otherwise tell the user that match not occur or may wrongly classify it.

The testing can be done by processing of one sample at time, but result can be shown one sample at time or many samples at a time; its all depends on machine user. Result of one sample test is shown in figure 7.7 and results of three samples (either correctly or non-correctly classified) are shown at time in figure 7.8.

The function not only reduces the risk of error that occurs due to wrong classification of wrong or un-available test vector but also calculate the % of correctly classified test vectors. The percentage is finding after processing of some predefined no of test samples i.e. after processing of every 6 test samples.

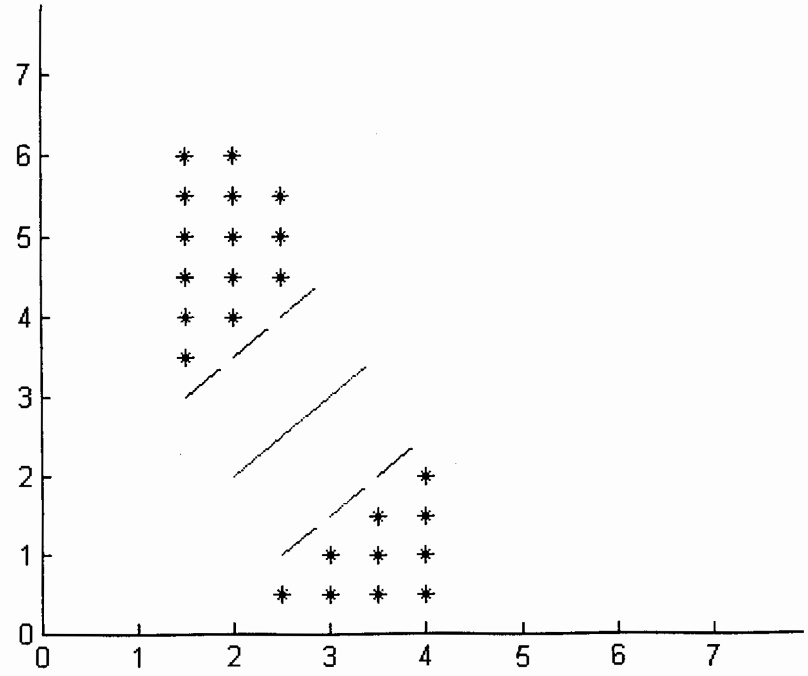


Figure 7.7: Testing of One Sample

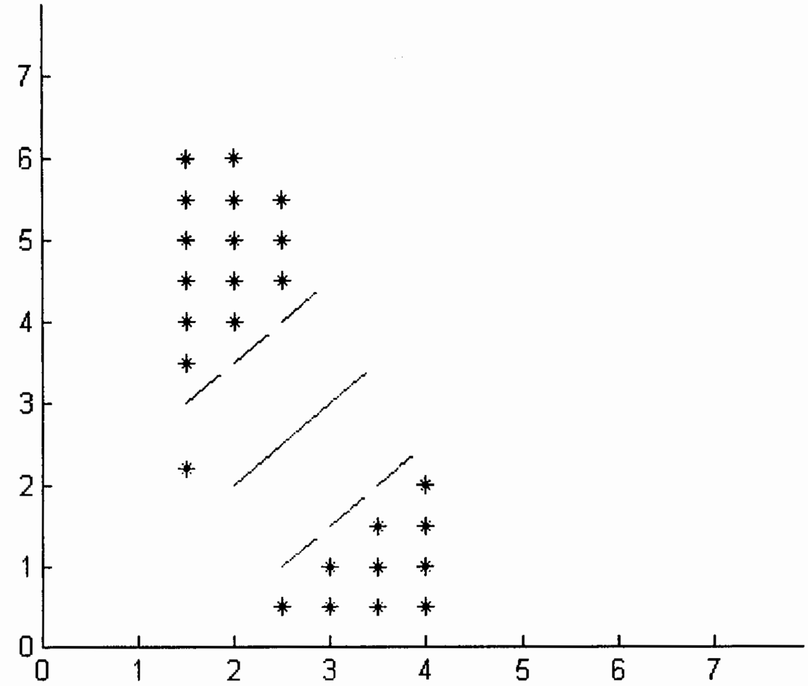
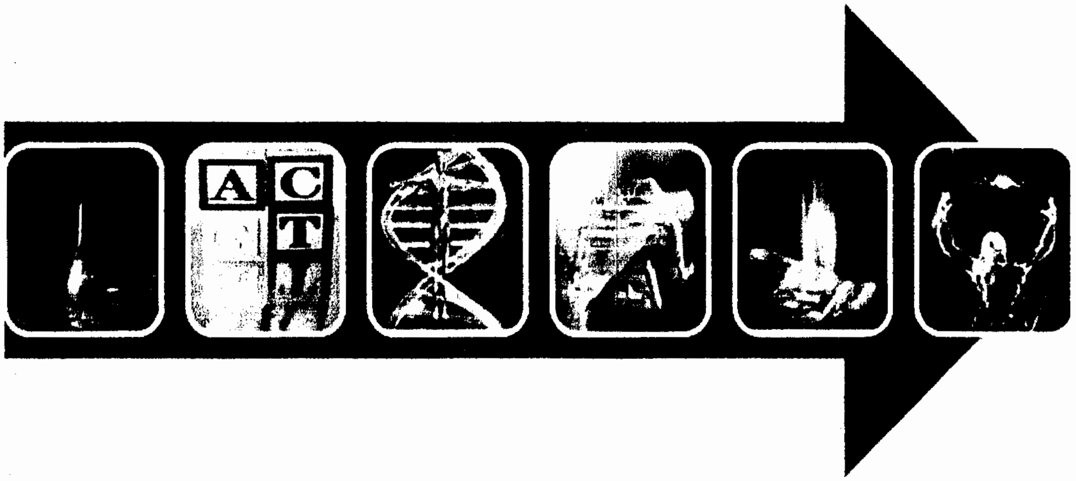


Figure 7.8: Testing of Three Samples

CHAPTER 6

TESTINGS & RESULTS



5. Testing & Results

Keeping in mind to fix the bugs, implemented machine is tested intended to verify that it satisfies its requirement and to discover the functional bugs.

System's testing is "Software quality Assurance" process of executing the implemented machine and comparing the experiential behavior to the preferred behavior.

5.1 Testing of Support Vector Machine

The testing of "Support Vector Machine" is undergone through all stages of black box testing and to extent white box testing. The machine is reviewed to see whether the objectives of the system are accomplished or not. A major factor considered during machine evaluation is to evaluate the machine with the perspective of "support vectors".

5.2 Results

During testing the sample vectors are classifies in either of two prescribed classes. The machine takes the name of sample vector from user and check to ensure that either it available on hard drive or not, if not then request to put the copy of test file of hard drive and then classify the trained sample by comparing the contents of test sample with all trained samples.

Machine maintained the databases of all sample vectors which are either trained & modified or tested. Three main results of the implemented intelligent machine are:

i. Feature Space can be increase and decrease as per user requisite
ii. Decision boundary is directly proportional to the sample vectors.
iii. Machine's performance is directly proportional to training vectors.

5.2.1 Analysis of Results

During training of the machine the training function call as many times as the input vectors are available that takes the name and label of file belongs to classes (CML. non-CML) and train it into respected class and insert the file into respected database table. Detail Analysis of results is accomplished subsequently.

5.2.1.1 Kernel Function Results

The implemented machine can use the linear kernel function that defines the SVM and satisfies the equation of line. Kernel function constructs a linear hyperplane as decision boundary in a prescribed feature space. During execution of kernel function feature space can be increase and decrease as per data requirement.

Feature space having decision boundary is shown in figure 8.1, and extended feature space having decision boundary before starting any trainings is represents in figure 8.2.

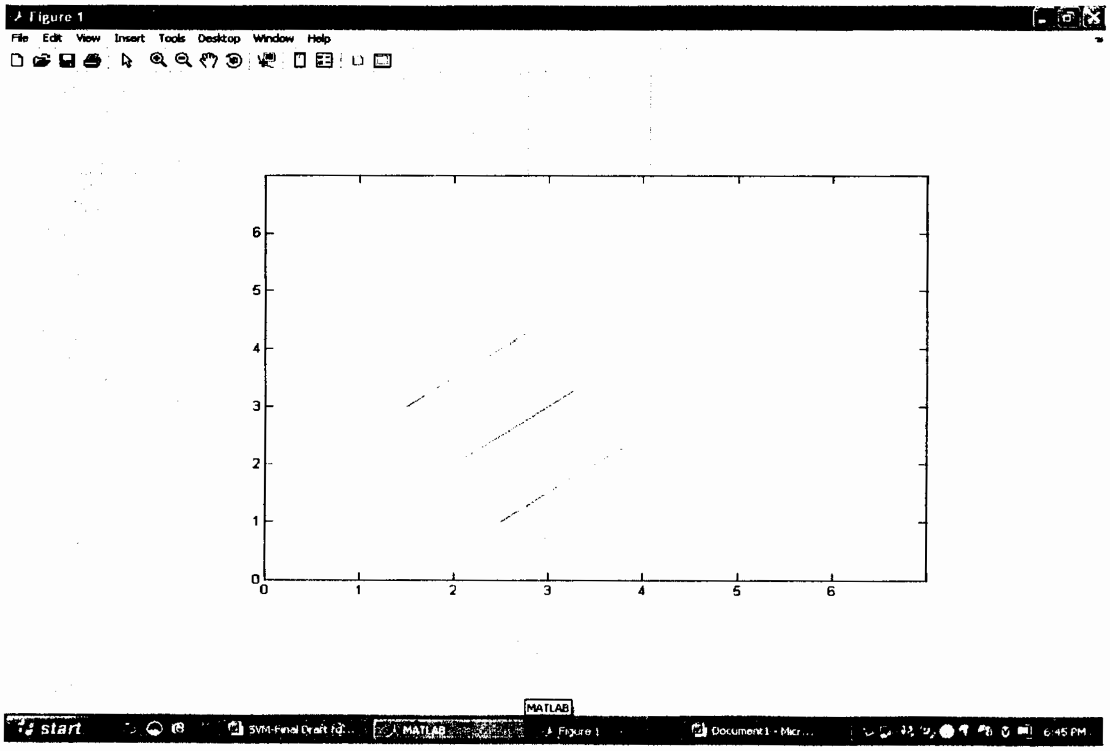


Figure 8.1: Feature Space with Decision Boundary before Training

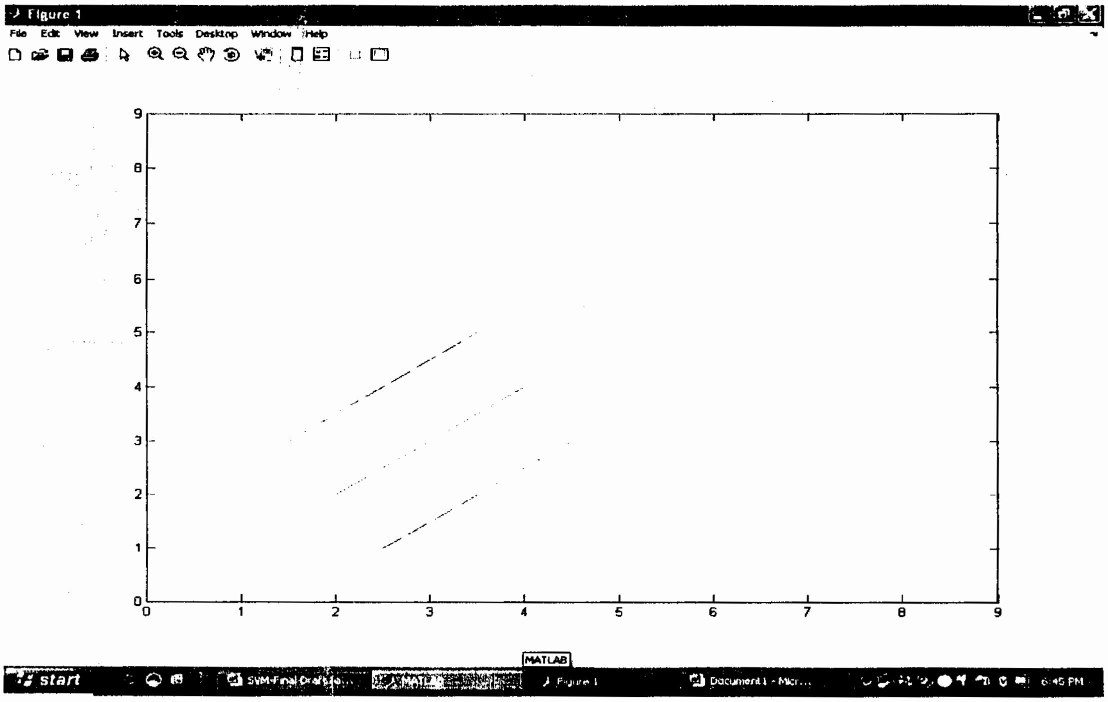
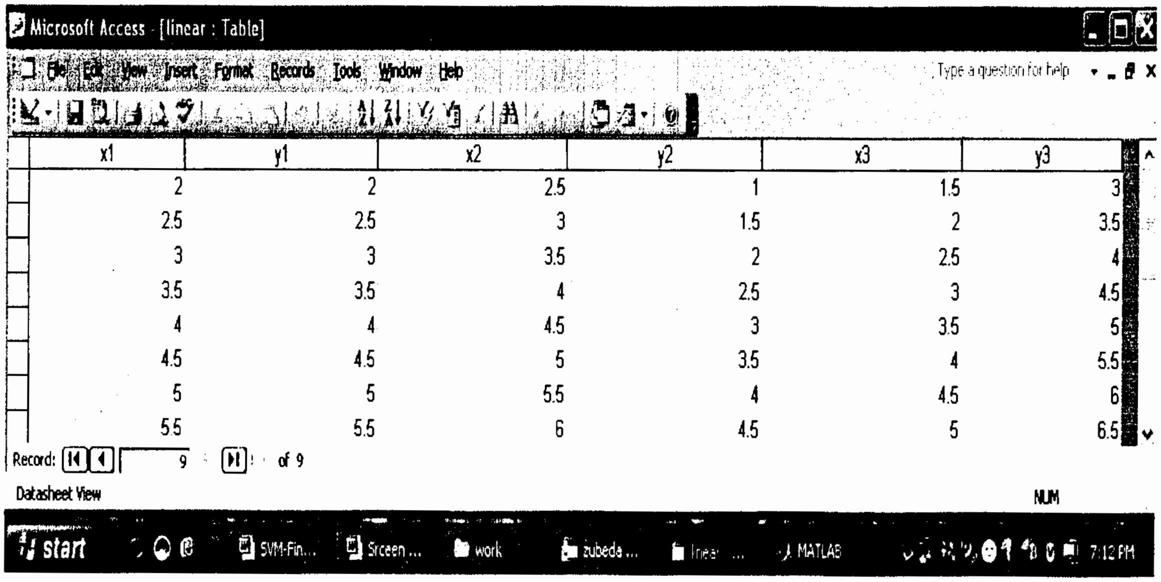


Figure 8.2: Extended Feature Space with Decision Boundary before Training

The function keeps records of all the points of linear hyperplane; either initial or extended that are being plotted to specify a decision boundary. These records are maintained by a database file. Sample of maintaining the record is shown in figure 8.3.



x1	y1	x2	y2	x3	y3
2	2	2.5	1	1.5	3
2.5	2.5	3	1.5	2	3.5
3	3	3.5	2	2.5	4
3.5	3.5	4	2.5	3	4.5
4	4	4.5	3	3.5	5
4.5	4.5	5	3.5	4	5.5
5	5	5.5	4	4.5	6
5.5	5.5	6	4.5	5	6.5

Figure 8.3: Database Table Associated with Kernel Function

5.2.1.2 Training Results

During training the machine train the sample input vector into two classes based on the label given to each input vector of both classes. The input vectors are the text file that contains the sample genetic sequences of different patients belongs to two classes of cancer i.e. CML & non-CML.

The machine takes the name and label of file belongs to classes from user and train it into respected class and insert the file into respected database table. When the function detects that text file is present of hard disk then train it successfully. If the input file is not present on the hard drive the firstly it demand to put the copy of file on hard drive and then train it. The trainings of input sample in a predefined decision boundary are proceeds one by one that is represented on figure 8.4.

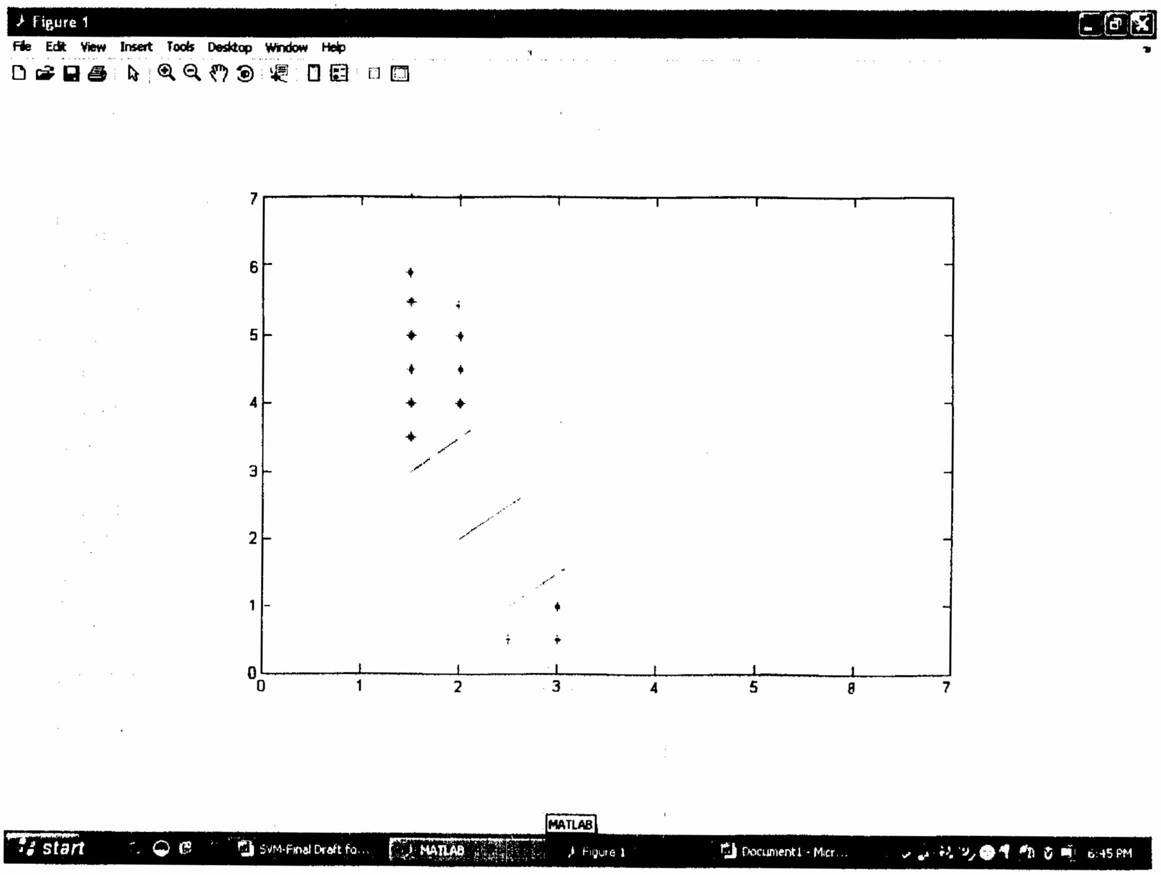


Figure 8.4: Training with Predefined Decision Boundary

During training the decision boundary will automatically increases with number of training input samples which is shown in figure 8.5.

The machine shows three dynamic properties during training; that are:

i. Feature space can be increased during training that is represented in figure 8.6.
ii. Decision boundary will automatically increases by increasing the number of sample
iii. During training the decision surface of each class will be increased and decreased as per data availability and training engineer requirement. Increasing and decreasing of decision surface are represented in figure 8.7 & 8.8 respectively.

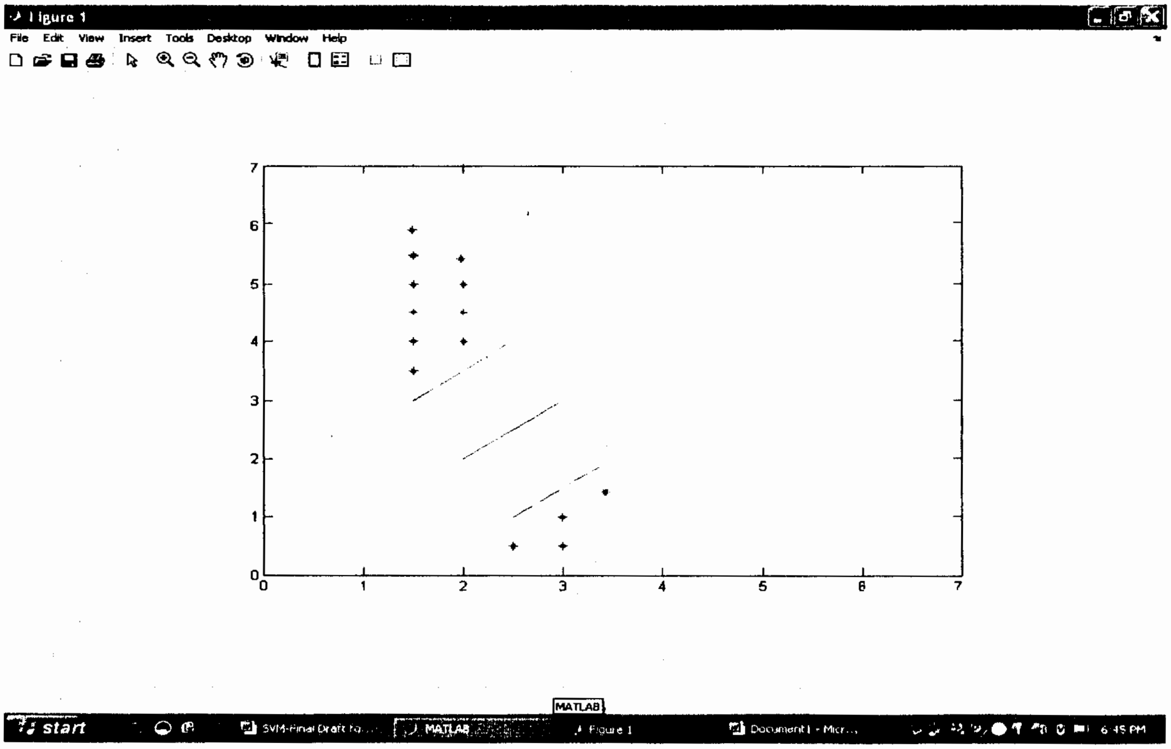


Figure 8.5: Increase of Decision Boundary During Training

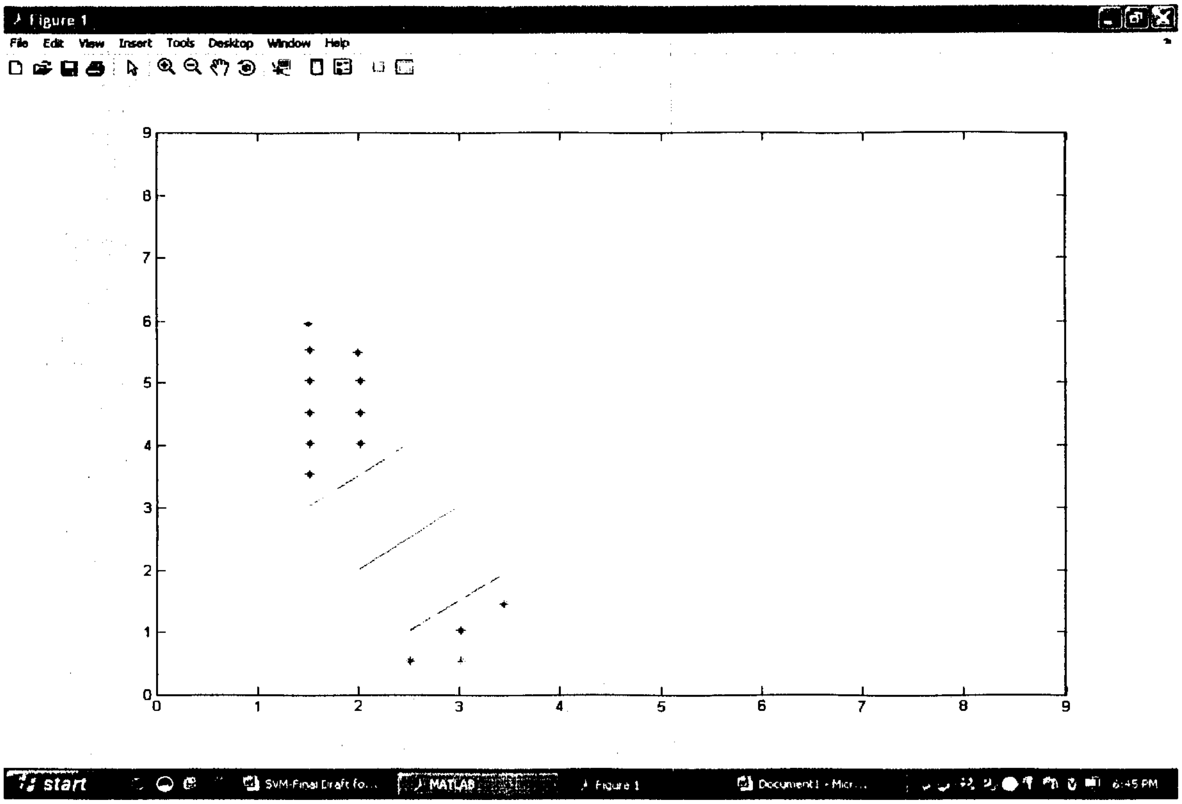


Figure 8.6: Increase of Feature Space During Training

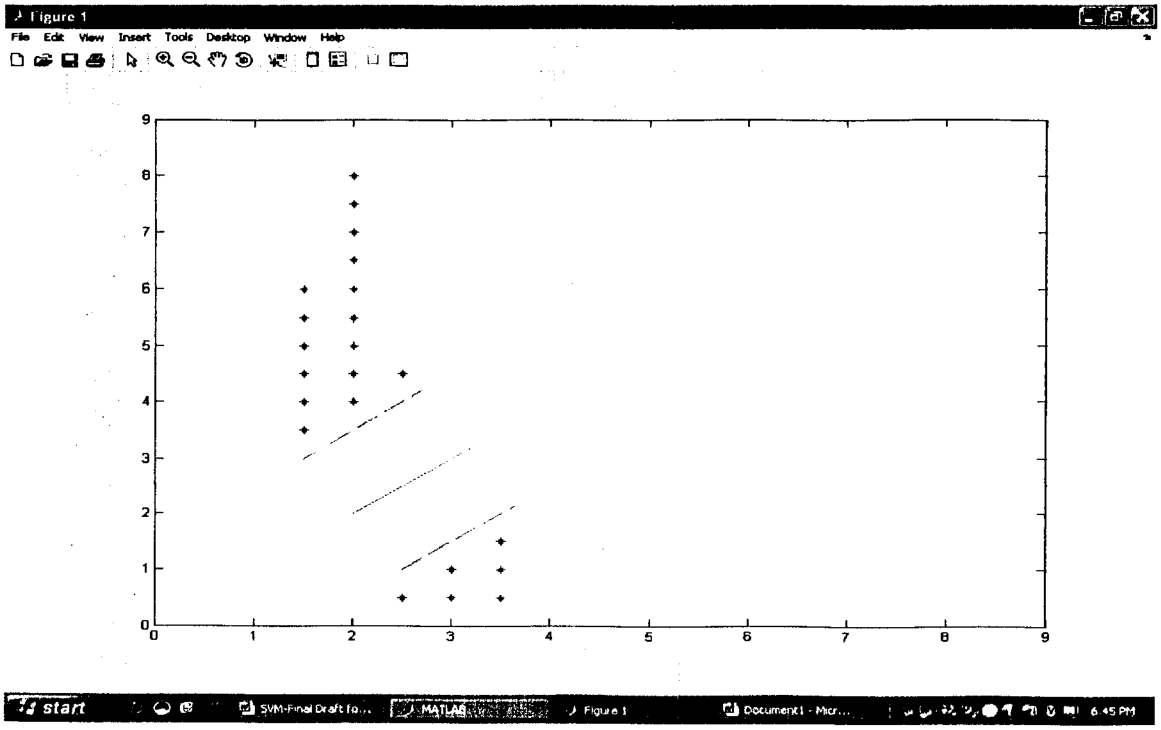


Figure 8.7: Increase of Decision Surface During Training

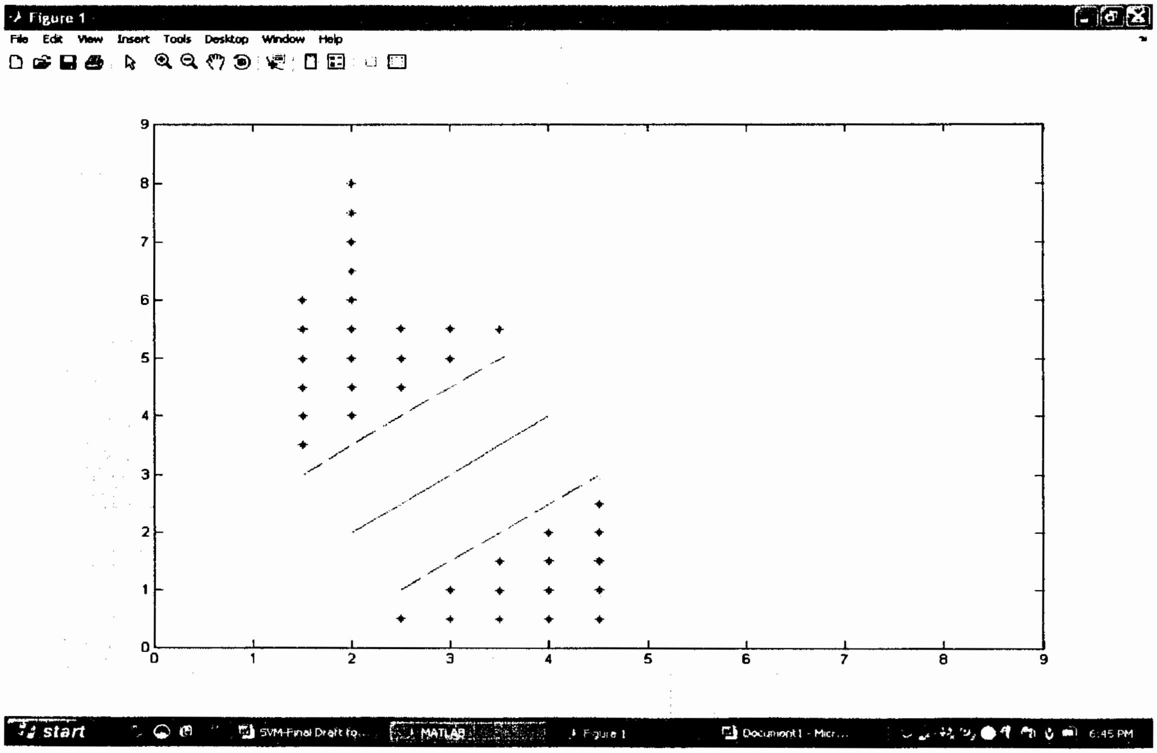
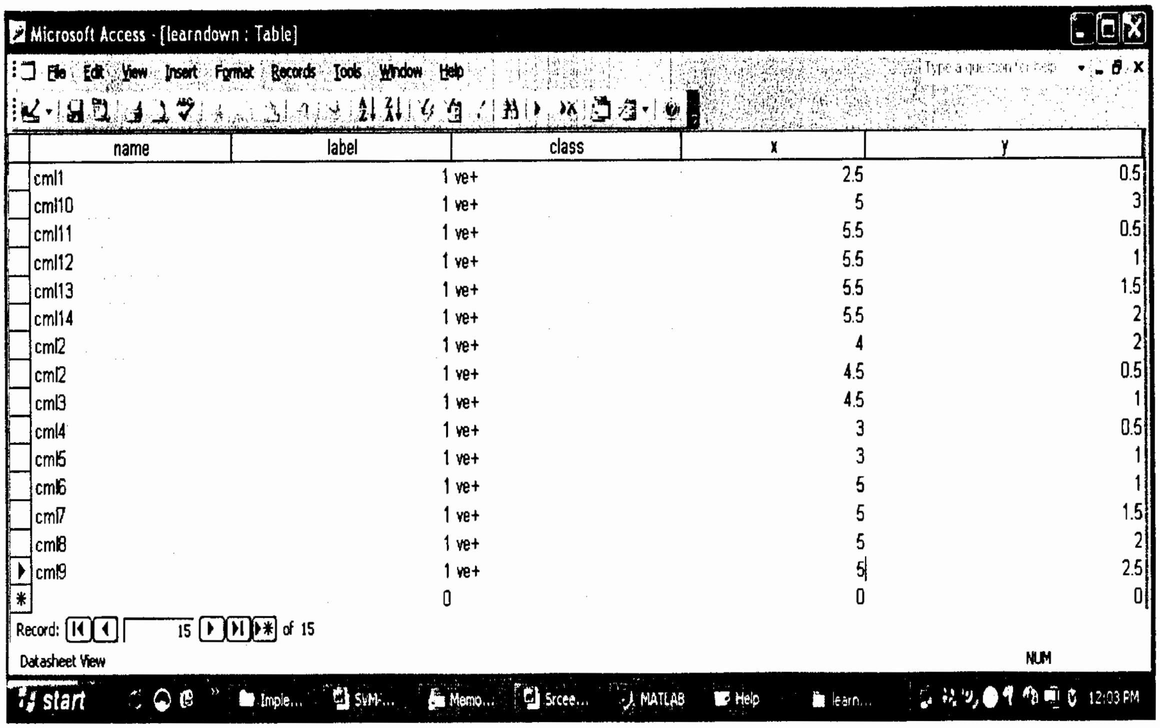


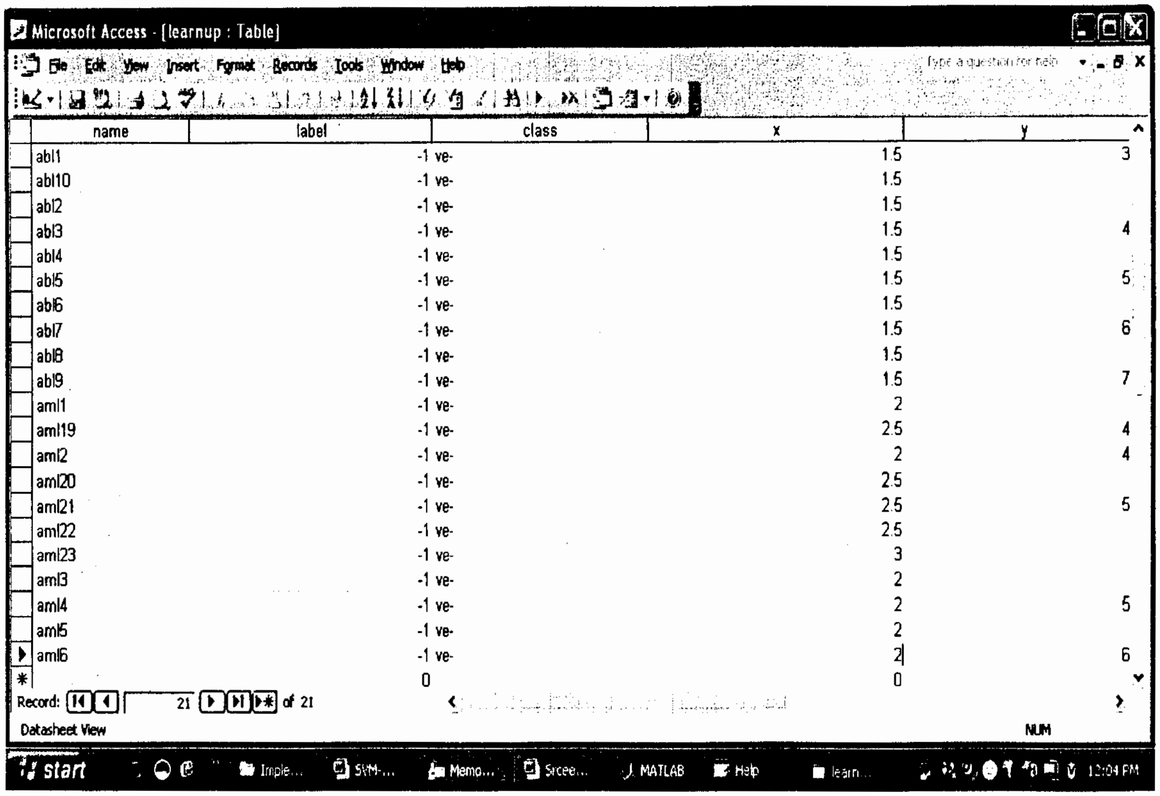
Figure 8.8: Decrease of Decision Surface During Training

During trainings of sample vectors the machine keeps the record of the entire training file in a correspondence database table. Database can separately maintain the training sample's records of both classes. Sample of these database files that machine will maintain are represented in figure 8.9 for CML class & in 8.10 for non-CML class. The samples of both classes are separated only on the basis of label that we give to each class sample at time of training. During testing, the machine also maintains the points of decision boundary in correspondence database table that is shown in figure 8.11.



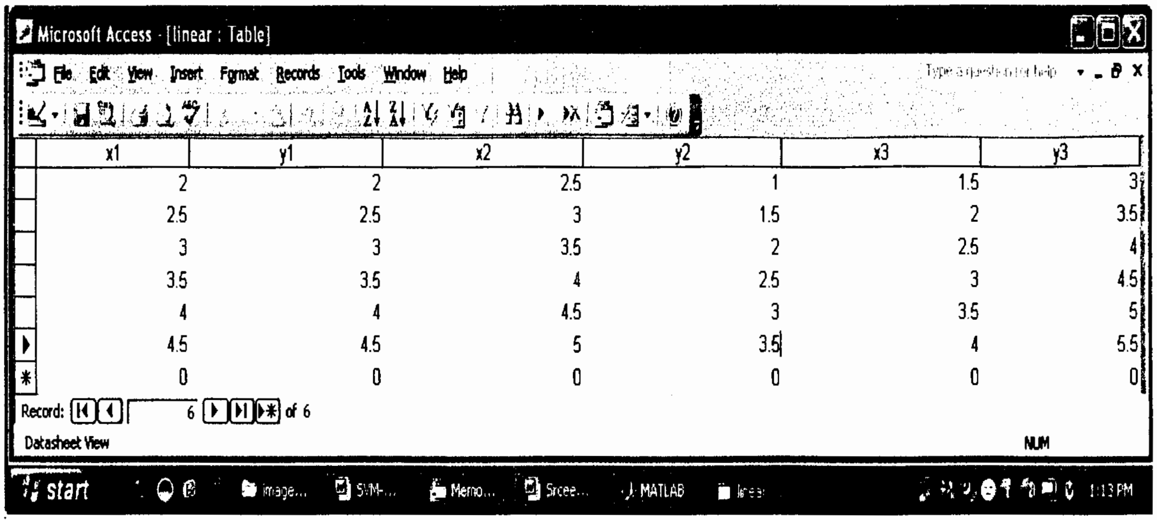
name	label	class	x	y
cml1	1	ve+	2.5	0.5
cml10	1	ve+	5	3
cml11	1	ve+	5.5	0.5
cml12	1	ve+	5.5	1
cml13	1	ve+	5.5	1.5
cml14	1	ve+	5.5	2
cml2	1	ve+	4	2
cml2	1	ve+	4.5	0.5
cml3	1	ve+	4.5	1
cml4	1	ve+	3	0.5
cml5	1	ve+	3	1
cml6	1	ve+	5	1
cml7	1	ve+	5	1.5
cml8	1	ve+	5	2
cml9	1	ve+	5	2.5
	0		0	0

Figure 8.9: Database Table Containing CML's Class Training Samples



name	label	class	x	y
abl1	-1	ve-	1.5	3
abl10	-1	ve-	1.5	
abl2	-1	ve-	1.5	
abl3	-1	ve-	1.5	4
abl4	-1	ve-	1.5	
abl5	-1	ve-	1.5	5
abl6	-1	ve-	1.5	
abl7	-1	ve-	1.5	6
abl8	-1	ve-	1.5	
abl9	-1	ve-	1.5	7
aml1	-1	ve-	2	
aml19	-1	ve-	2.5	4
aml2	-1	ve-	2	4
aml20	-1	ve-	2.5	
aml21	-1	ve-	2.5	5
aml22	-1	ve-	2.5	
aml23	-1	ve-	3	
aml3	-1	ve-	2	
aml4	-1	ve-	2	5
aml5	-1	ve-	2	
aml6	-1	ve-	2	6
	0		0	

Figure 8.10: Database Table Containing non-CML's Class Training Samples


Microsoft Access - [linear : Table]
File Edit View Insert Format Records Tools Window Help
Record:
of 6
Datasheet View
NUM
start
MATLAB
1:13 PM

x1	y1	x2	y2	x3	y3
2	2	2.5	1	1.5	3
2.5	2.5	3	1.5	2	3.5
3	3	3.5	2	2.5	4
3.5	3.5	4	2.5	3	4.5
4	4	4.5	3	3.5	5
4.5	4.5	5	3.5	4	5.5
0	0	0	0	0	0

Figure 8.11: Decision Boundary in Correspondence Database Table

5.2.1.3 Modification Results

The machine reduces the risk of error that occurs due to training of input vector that may contain wrong or outdated information. The machine has a capability to the wrong trained file from the training system.

The modifications of one CML's class input sample from the result shown in figure 8.8 is represented in figure 8.12. The modified entry also deleted form the correspondence database table. i.e. In figure 8.8 and figure 8.9 there are 15 input sample of CML but after modification; both figure shows 14 entries for +ve label class. Correspondence database table is shown in figure 8.13.

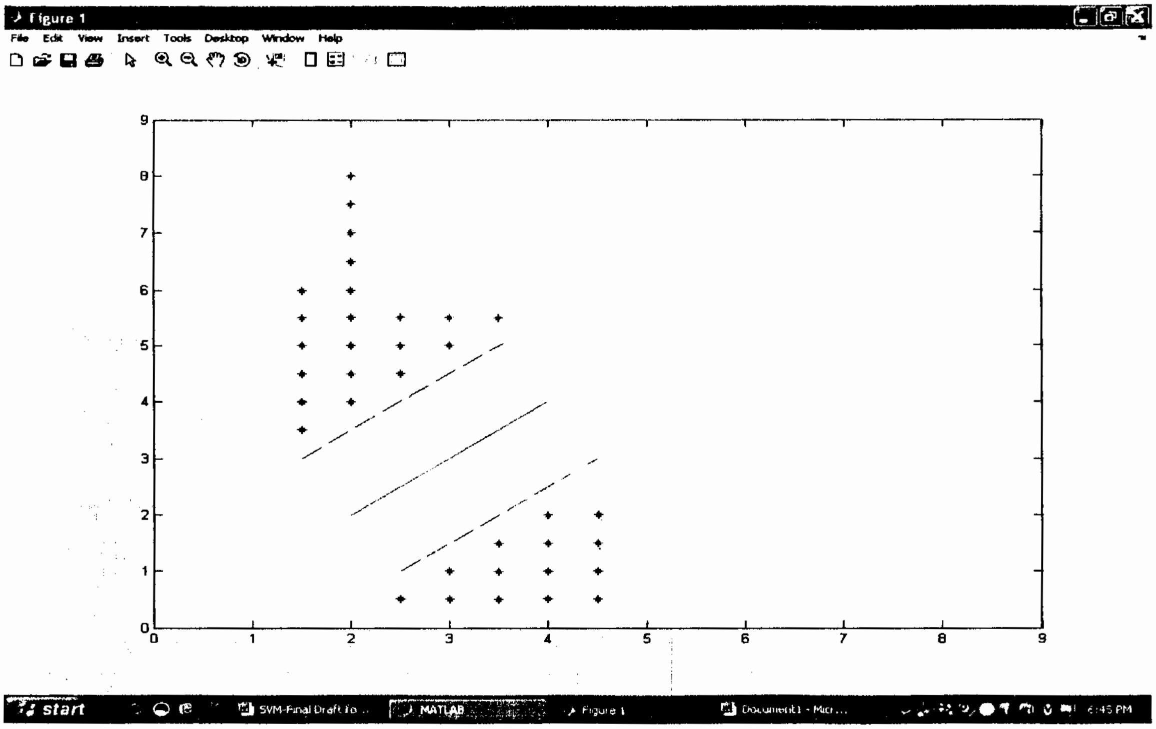
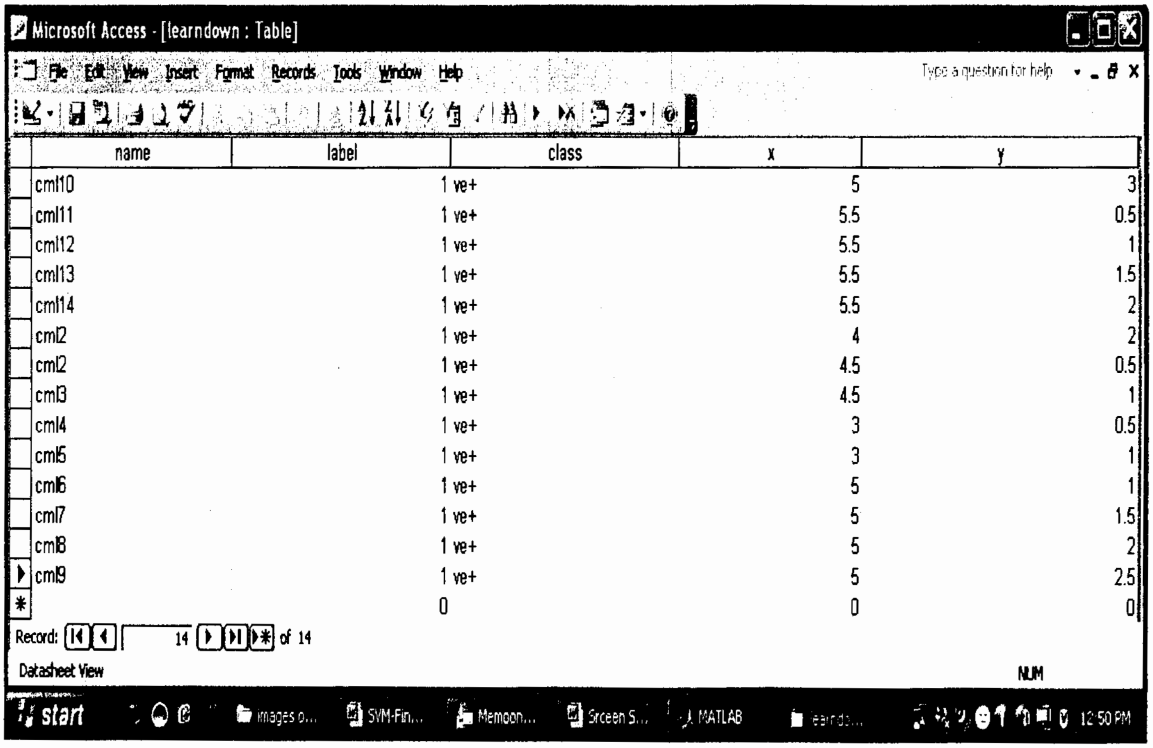


Figure 8.12: Modification of one Sample from CML Class



name	label	class	x	y
cml10	1	ve+	5	3
cml11	1	ve+	5.5	0.5
cml12	1	ve+	5.5	1
cml13	1	ve+	5.5	1.5
cml14	1	ve+	5.5	2
cml2	1	ve+	4	2
cml2	1	ve+	4.5	0.5
cml3	1	ve+	4.5	1
cml4	1	ve+	3	0.5
cml5	1	ve+	3	1
cml6	1	ve+	5	1
cml7	1	ve+	5	1.5
cml8	1	ve+	5	2
cml9	1	ve+	5	2.5
	0		0	0

Record: 14 of 14

Datasheet View

NUM

Figure 8.13: Database Table Representing Modification of one Sample from CML Class

The modifications of one non-CML's class input sample from the result shown in figure 8.8 is represented in figure 8.14. The modified entry also deleted form the correspondence database table. i.e. In figure 8.8 and figure 8.10 there are 21 input sample of non-CML but after modification; both figure shows 20 entries for -ve label class. Correspondence database table is shown in figure 8.15.

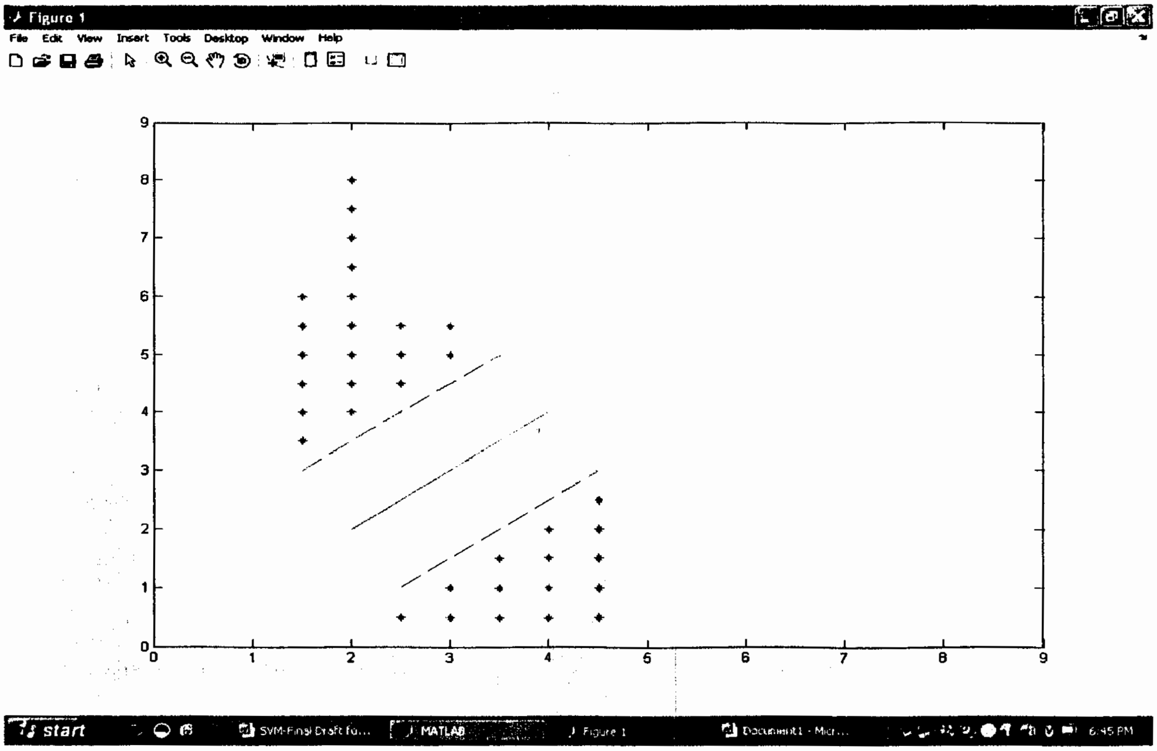
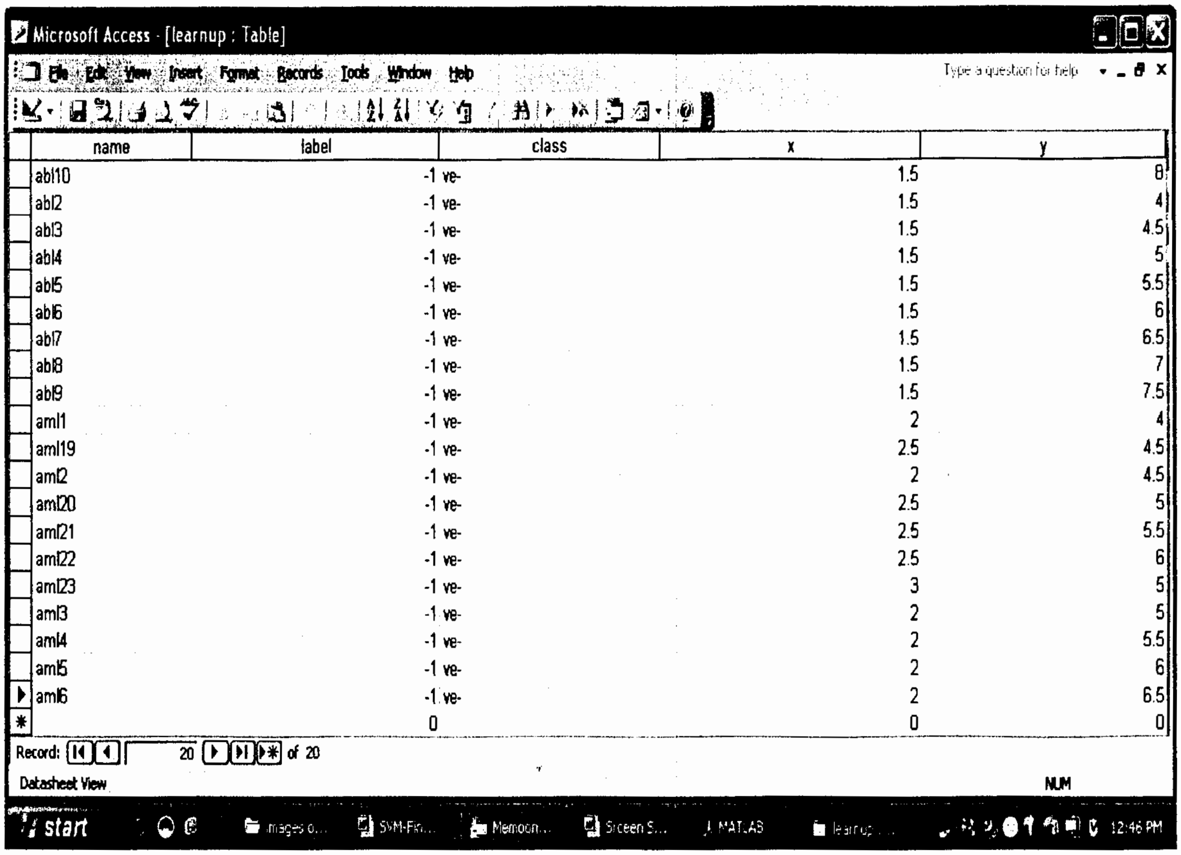


Figure 8.14: Modification of one Sample from non-CML Class



name	label	class	x	y
abl10	-1	ve-	1.5	8
abl2	-1	ve-	1.5	4
abl3	-1	ve-	1.5	4.5
abl4	-1	ve-	1.5	5
abl5	-1	ve-	1.5	5.5
abl6	-1	ve-	1.5	6
abl7	-1	ve-	1.5	6.5
abl8	-1	ve-	1.5	7
abl9	-1	ve-	1.5	7.5
aml1	-1	ve-	2	4
aml19	-1	ve-	2.5	4.5
aml2	-1	ve-	2	4.5
aml20	-1	ve-	2.5	5
aml21	-1	ve-	2.5	5.5
aml22	-1	ve-	2.5	6
aml23	-1	ve-	3	5
aml3	-1	ve-	2	5
aml4	-1	ve-	2	5.5
aml5	-1	ve-	2	6
aml6	-1	ve-	2	6.5
	0		0	0

Figure 8.15: Database Table Representing Modification of one Sample from CML Class

During the process of deletion; decision boundary is automatically decreases by decreasing the sample vectors. Modifications of samples are processed one by one and decrease the decision boundary. After modification of six samples from the result shown in figure 8.8 that belongs to both classes machine automatically decreases the decision boundary that is represented in figure 8.16. Machine keeps the record of decreasing points of decision boundary in correspondence database table. i.e. Figure 8.11 containing the record of decision boundary of figure 8.8. After deleting the training sample form figure 8.8 the decision boundary decrease, the record of point of decision boundary are shown in figure 8.17.

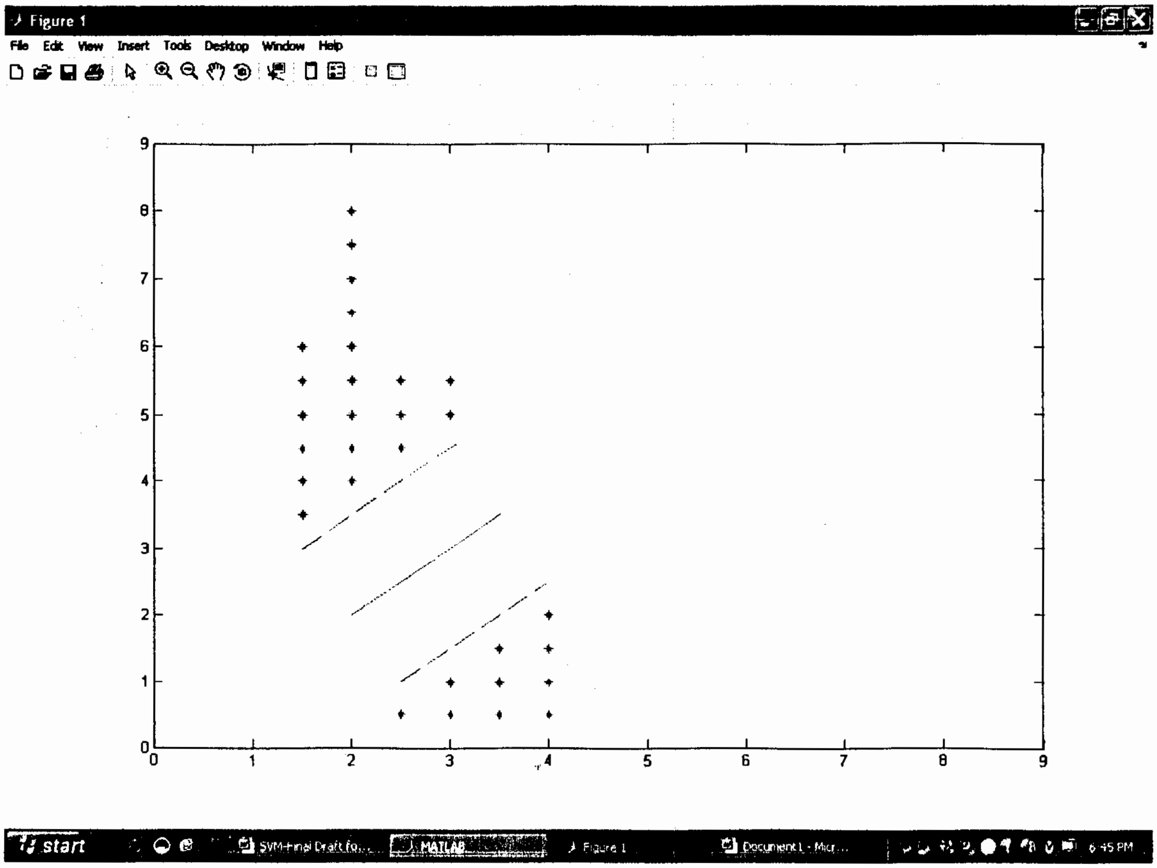
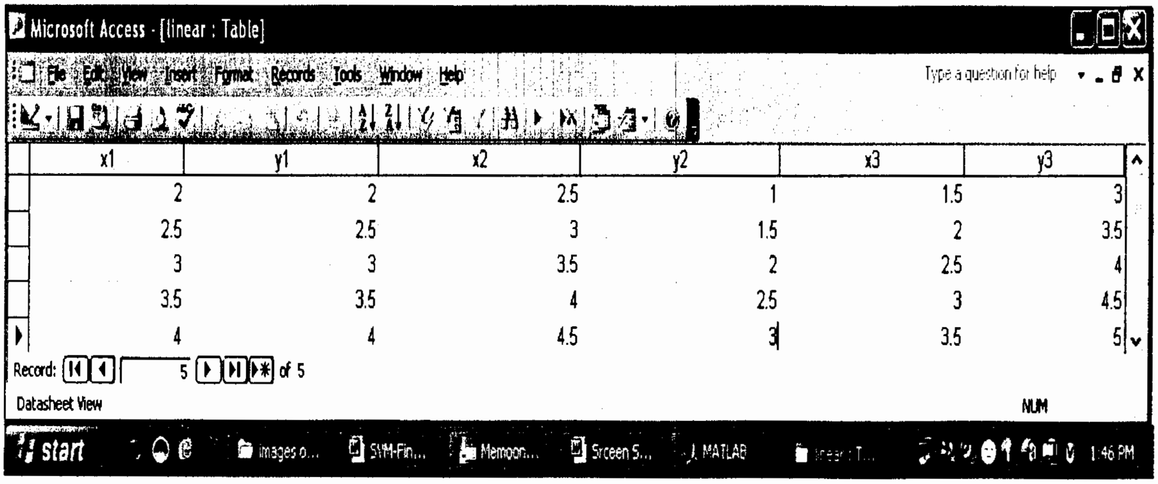


Figure 8.16: Decreasing of Decision Boundary after Modification


Microsoft Access - [linear : Table]
File Edit View Insert Format Records Tools Window Help
Type a question for help
Record: 5 of 5
Datasheet View
NUM
start
MATLAB
1:46 PM

x1	y1	x2	y2	x3	y3
2	2	2.5	1	1.5	3
2.5	2.5	3	1.5	2	3.5
3	3	3.5	2	2.5	4
3.5	3.5	4	2.5	3	4.5
4	4	4.5	3	3.5	5

Figure 8.17: Table keeping the Record of Decision Boundary after Modification

5.2.1.4 Testing Results

The machine performs the following functions during testing:

i. Firstly machine takes the name of sample vector from user
ii. Check the availability of test file on hard drive
iii. Request the test engineer to insert the copy of test file o hard drive if check of availability of test file is false.
iv. Check the availability of test file in trained sample
v. Gives the message to test engineer either the test file matched or not matched
vi. Gives the % results after testing of six test sample

If the machine doesn't detect the test file with that name then it firstly requested to place a copy of test sample's text file on the hard drive. But if the copy of test file is detected then compares the contents of that file with all the trained vectors and classifies it into respected class and gives the message to user that file is matched. Otherwise tell the user that match not occur or may wrongly classify it.

The machine can precedes the testing by one sample at time, but result can be shown one sample at time or many samples at a time; its all depends on machine user. Result of one correct sample test that belongs to CML class is shown in figure 8.18. Result of one correct sample test that belongs to non-CML class is shown in figure 8.19 and Result of one misclassification of sample test that belongs to neither of both classes is shown in figure 8.20. Results of six samples (either correctly or incorrectly classified) are shown at time in figure 8.21.

During testing of the machine keeps the record of all tested samples in correspondence database table as shown in figure 8.22.

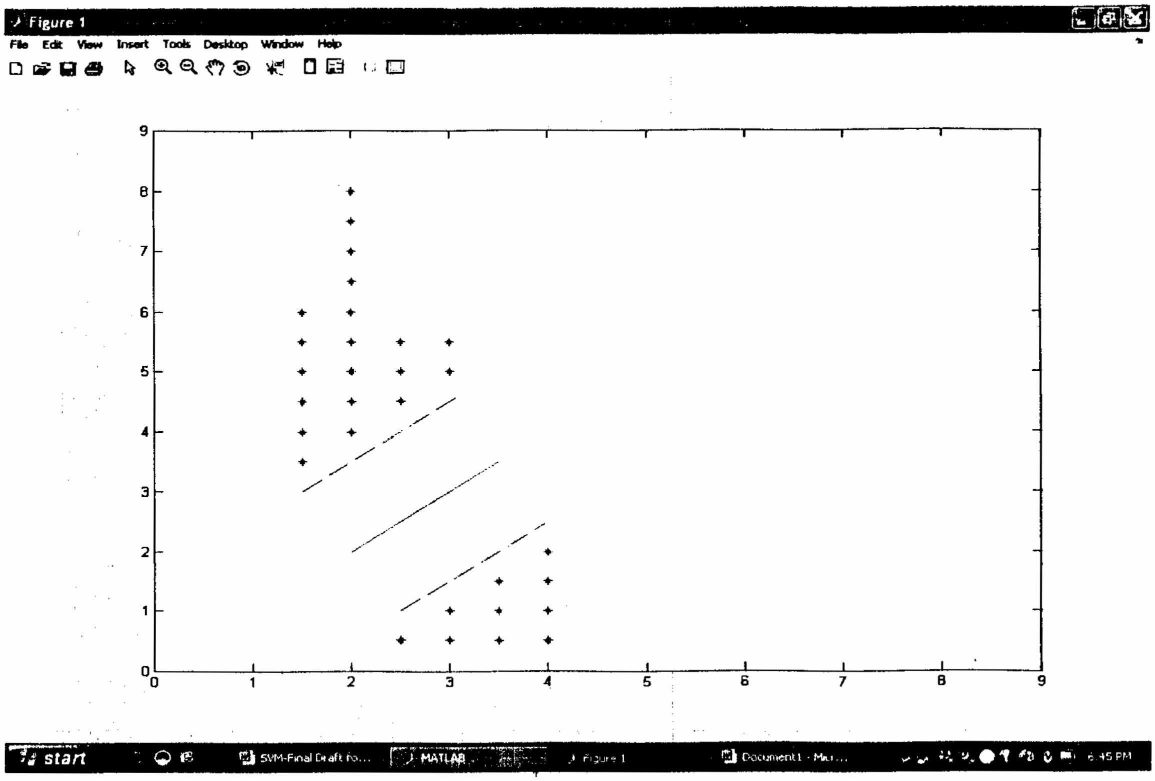


Figure 8.18: Correct Classification of one Test Sample Belongs to CML Class

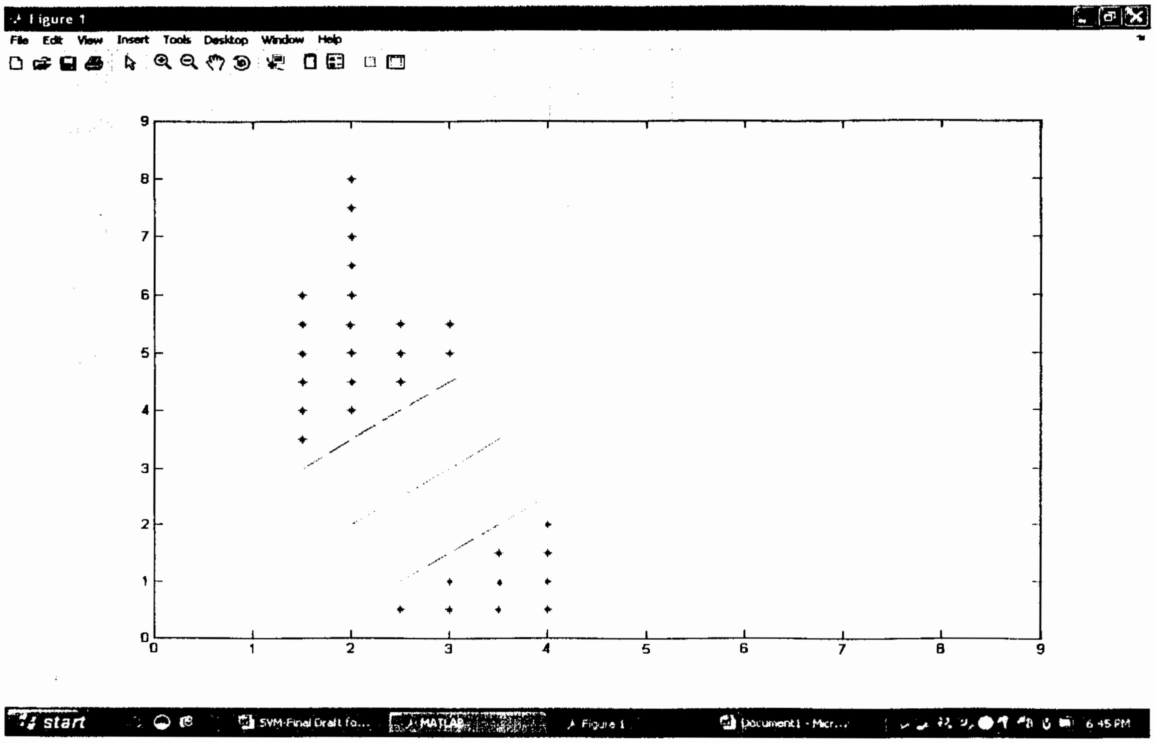


Figure 8.19 Correct Classification of one Test Sample Belongs to non-CML Class

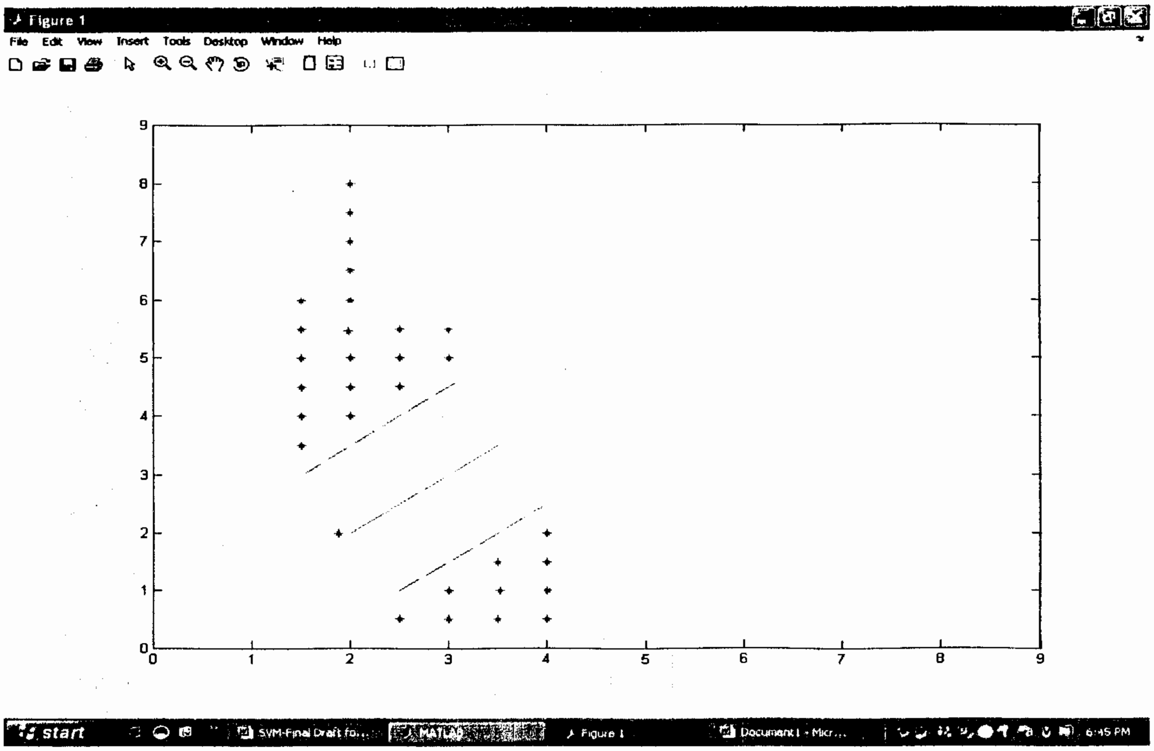


Figure 8.20: Correct Classification of one Test Sample Belongs to Neither of Classes

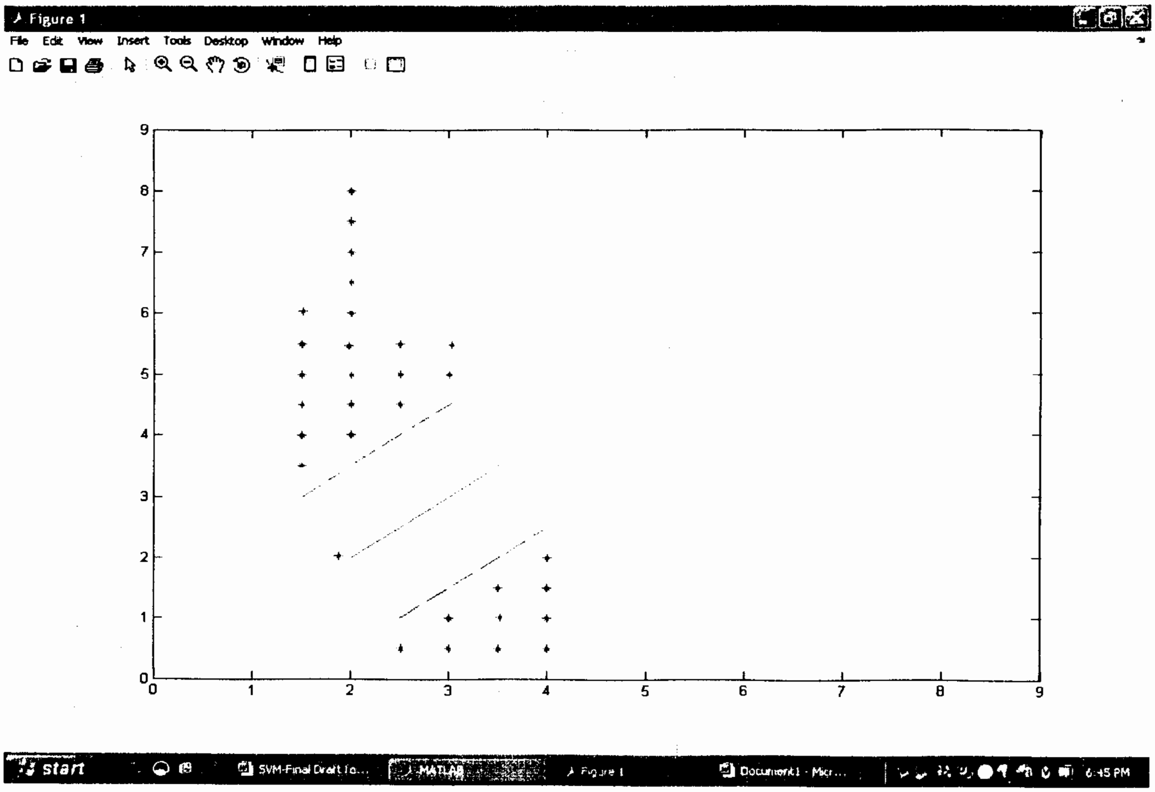
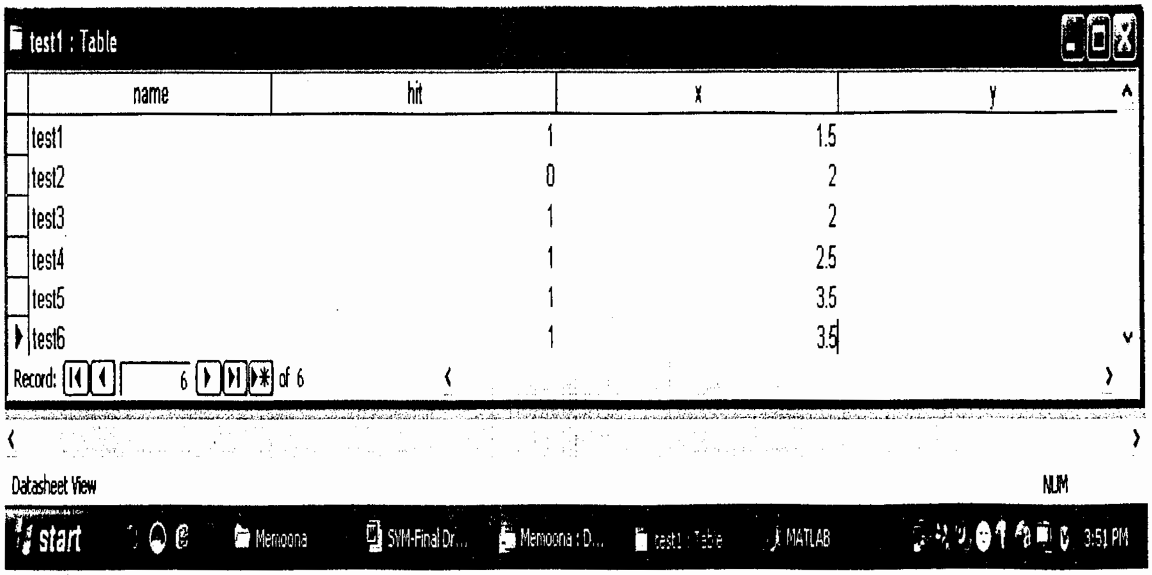


Figure 8.21: Classification of Six Test Sample Belongs to Both of Classes



test1 : Table

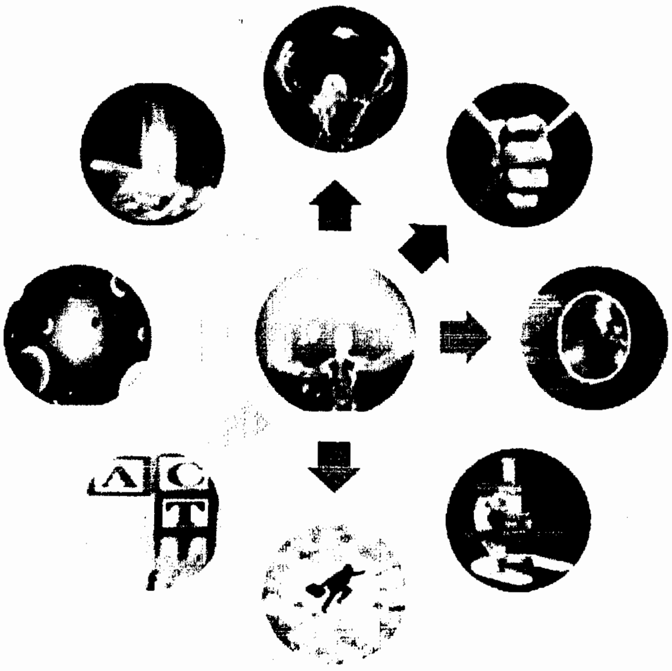
name	hit	x	y
test1	1	1.5	
test2	0	2	
test3	1	2	
test4	1	2.5	
test5	1	3.5	
test6	1	3.5	

Record: 6 of 6

Figure 8.22: Database Table Maintaining the Record of Test Samples

CHAPTER 6

CONCLUSION & FUTUR ENHANCEMENT



6. Conclusion & Future Enhancement

Recent years have seen dramatic and sustained growth in the amount of genomic data being generated, including in late 1999 the first complete sequence of a human chromosome. The challenge faced by biological scientists is to make sense of this vast amount of accumulated and accumulating data. Fortunately, numerous techniques are provided as resources that analyze this data and attempt to understand it.

6.1 Conclusion

In the current research we use the SV machine because in computational biology researcher advocates two main incentives for the use of SVMs. First, in this biological problem high-dimensional, noisy data is involved, for which SVMs are known to perform well as compare to other machine learning methods. Second, non-vector inputs can easily handled by kernel methods like the SVM.

SVM can easily solve the problem of analyzing the genomic data and amalgamate the primary databases with software in a flexible and robust way. One must consider that the performance of the implemented system is directly affected with the number of training vector.

The machine not only reduces the risk of error that occurs due to wrong training of input vector but also reduces the risk of error that occurs due to wrong classification of wrong or un-available test vector

6.2 Future Enhancement

In future all the architecture proposed system will try to soft code implemented because in current research first two modules are hard code implemented.

In future we try to make a intelligent machine the correctly classify the all types of blood cancer, and this will done by comprehensive understanding of the genetic alterations present in all tumors.

Different type of classification tools will use and results will compare to make the best intelligent machine that will detect and diagnose the tumor very sharply without any error.

REFERENCES

REFERNCES

- Cai, J.; Dayanik, A.; Yu, H.; Hasan, N.; Terauchi, T. 7 & Noble, W. Grundy: **"Classification of Cancer Tissue Types by Support Vector Machines Using Microarray Gene Expression Data."** Department of Medical Informatics, Department of Computer Science, Columbia University, New York and Department of Computer Science, Rutgers University, Piscataway, 2000.
- Cho, S. Bae & Won, H. Hee: **"Machine Learning in DNA Microarray Analysis for Cancer Classification."** Conferences in Research and Practice in Information Technology. Vol.19, 2003.
- Fujarewicz, K.; Swierniak, A. Gliwice; Poland, B. Jarzab; Wiench, M. Silesian & Kimmel, M.: **"Using Support Vector Machines for analysis of gene expression data from DNA microarrays."** Centre of Oncology - Institute of Oncology, Poland, University of Technology and Rice University, Houston, Texas, 2003.
- http://www.genome.gov/11508982, 2007
- http://www.kernel-machines.org, 2007
- http://www.ncbi.nlm.nih.gov, 2007
- http://www.wikipedia.org, 2007
- Komura, D.; Nakamura, H.; Tsutsumi, S.; Aburatani, H.; Ihara, S.: **"Characteristics of Support Vector Machines in Gene Expression Analysis."** Genome Informatics 13, pp. 264-265, 2002.
- Kurzrock, R. Hagop M. Kantarjian, Brian J. Druker, and Moshe T.: **"Philadelphia Chromosome–Positive Leukemias: From Basic Mechanisms to Molecular Therapeutics"**, Ann Intern Med.; Vol. 138, pp. 819-830, 2003.
- Laurent, E.; Talpaz,M.; Kantarjian, H. & Kurzrock, R.: **"The BCR Gene and Philadelphia Chromosome-positive Leukemogenesis."** Cancer Research, 61, 2343–2355, March 15, 2001.
- Michael, P. S. Brown; Grundy, N. William; Lin, D.; Cristianinit, N.; Sugnet, C.; Ares, M. & Hausslerz, D.: **"Support Vector Machine Classification of**

Microarray Gene Expression Data." University of California, Santa Cruz, University of Bristol, UK, 12 June, 1999.

- Yang, Z. Rong: **"Biological applications of support vector machines."** Briefing in Bioinformatics. Vol. 5, No. 4, pp. 328–338, December 2004.