CHAPTER 4

EFFICIENT CLASSIFICATION WITH MULTICLASS SVM ENSEMBLES

4.1 INTRODUCTION

Recently, the utilization of data mining is vital in a wide range of fields for the knowledge extraction process. The gene data is in the expression levels of thousands of genes for a small amount of samples. Acquiring knowledge is the intricacy in such types of gene data. Extremely huge quantity of data and small amount of knowledge extraction competence increases the interest in the field of data mining states Satchidananda & Cho (2008). Data Mining or the proficient innovation of significant information from a huge collection of data has a goal to discover knowledge out of data and present it in a form that is easily comprehensible to humans mentions Nevine & Michael (2007). While the DNA microarray technology considerably speed up the process of finding the utility of genes, the amount of data generated by this technology also poses a challenge for the biologists to perform the analysis. (Hsinchun 2008) narrates in microarray data analysis, the process of information retrieval system comprises diagnosis of disease, classifying disease, and retrieving information which is helpful to provide necessary treatments.

New method to analyze gene data consists of both classification of the tissue samples, and an exploration of the data for mislabeled or questionable tissue results. Microarray expression experiments allow the recording of expression levels of thousands of genes simultaneously. These experiments primarily consist of either monitoring each gene multiple times under many conditions or alternately evaluating each gene in a single environment but in different types of tissues, especially cancerous tissues. In recent years several methods have been developed
for performing gene expression experiments. Measurements from these experiments can give expression levels for genes in tissue or cell samples. Microarrays, also known as gene chips or DNA chips, provide a convenient way of obtaining gene expression levels for a large number of genes simultaneously. Each spot on a microarray chip contains the clone of a gene from a tissue sample. For example, gene expression data have been used to obtain results in the classifications of lymphoma, leukemia, breast cancer, and liver cancer.

The gene classification is one of the major processes in microarray gene expression analysis explains Changjing & Qiang (2005) since it is a basis for prediction of the functions of unknown genes predicts Seungchan et al (2006). Such classification measures usually have the stages gene selection or dimension reduction, where a small amount of gene components are created from a vast number of genes and classification, where the samples are classified into groups by applying standard statistical models on the gene components illustrates Slavkov et al (2005). Normally, microarray datasets is of thousands of genes in a few dozens of samples, thus very accurate arrangement of tissue samples in such high dimensional problems is a complicated task says (Jae-Woo 2007). This chapter presents an effective microarray gene data classification with the aid of PCA and SVM. Here dimensionality diminution process is carried out in order to shrink the microarray data without losing information with the aid of PCA. An attempt was made to present a broad assessment of more than a single class (Multiclass) with the aid of this SVM ensemble. The results presented are based on experiments performed on the evaluation set.

4.2 MULTICLASS CLASSIFICATION

Classification and prediction are two forms of data analysis that can be used to extract models describing important data classes or to predict future data trends. Such analysis can help provide us with a better understanding of the data at large.
Data classification is a two-step process, in the first step; a classifier is built describing a predetermined set of data classes or concepts. This is the learning step (or training phase), where a classification algorithm builds the classifier by analyzing or “learning from” a training set made up of database tuples and their associated class labels. A tuple, $X$, is represented by an n-dimensional attribute vector, $X = (x_1, x_2, \ldots, x_n)$, depicting n measurements made on the tuple from n database attributes, respectively, $A_1, A_2, \ldots, A_n$. Each tuple, $X$, is assumed to belong to a predefined class as determined by another database attribute called the class label attribute. The class label attribute is discrete-valued and unordered. It is categorical in that each value serves as a category or class. The individual tuples making up the training set are referred to as training tuples and are selected from the database under analysis. Because the class label of each training tuple is provided, this step is also known as supervised learning (i.e., the learning of the classifier is “supervised” in that it is told to which class each training tuple belongs). It contrasts with unsupervised learning (or clustering), in which the class label of each training tuple is not known, and the number or set of classes to be learned may not be known in advance.

The process of information retrieval can operate on one of two broad types of information sources: structured and unstructured. Structured sources are exemplified by records within a database and standardized forms, where certain information must appear in specific locations. Unstructured sources are common in most fields, with examples including books, articles, and most other informal means of human communication. Performing retrieval on structured data sources is usually a simple task. A search is executed for the contents of a combination of predefined fields, for example, retrieving records from a database based on the ranges of a certain field. Each training point belongs to one of N different classes. The goal is to construct a function which, given a new data point, will correctly predict the class to which the new point belongs mentions Ofer & Ohad (2010).
4.2.1 One against the Rest Approach

This method is also called winner-take-all classification or one against all approach. Suppose the dataset is to be classified into M classes. Therefore, M binary SVM classifiers may be created where each classifier is trained to distinguish one class from the remaining M-1 classes. For example, class one binary classifier is designed to discriminate between class one data vectors and the data vectors of the remaining classes. During the testing or application phase, data vectors are classified by finding margin from the linear separating hyperplane. The final output is the class that corresponds to the SVM with the largest margin. This multiclass method has an advantage in the sense that the number of binary classifiers to construct equals the number of classes reveals (Mahesh 2008). However, there are some drawbacks. First, during the training phase, the memory requirement is very high and amounts to at the square of the total number of training samples. This may cause problems for large training data sets and computer memory problems. Second, suppose there are M classes and each has an equal number of training samples. During the training phase, the ratio of training samples of one class to rest of the classes will be 1: (M –1). In practice, on-vs-rest classification is usually preferred, since the results are mostly similar, but the runtime is significantly less points out Yi & Yuan (2006).

4.2.2 One against One approach

In this method, SVM classifiers for all possible pairs of classes are created. Therefore, for M classes, there will be binary classifiers. The output from each classifier in the form of a class label is obtained. The class label that occur the most is assigned to that point in the data vector. In case of a tie, a tie-breaking strategy may be adopted. A common tie-breaking strategy is randomly select one of the class labels that are tied. The number of classifiers created by this method is generally much larger than the previous method. However, the number of training data vectors required for each classifier is much smaller. Therefore, this method is
considered more symmetric than the one against the rest method. Moreover, the memory required to create the kernel matrix is much smaller. However, the main disadvantage is the increase in the number of classifiers as the number of class increases. For example, for 7 classes of interest, 21 classifiers will be created.

### 4.2.3 Decision Directed Acyclic Graph based approach

This is based on the Decision Directed Acyclic Graph (DDAG) structure that forms a tree-like structure. The DDAG method in essence is similar to pairwise classification such that, for an $M$ class classification problem, the number of binary classifiers is equal to $0.5^*M(M-1)$ and each classifier is trained to classify two classes of interest. Each classifier is treated as a node in the graph structure. Nodes in DDAG are organized in a triangle with the single root node at the top and increasing thereafter in an increment of one in each layer until the last layer that will have $M$ nodes. The DDAG evaluates an input vector $\mathbf{x}$ starting at the root node and moves to the next layer based on the output values. For instance, it exits to the left edge if the output from the binary classifier is negative, and it exits to the right edge if the output from the binary classifier is positive. The binary classifier of the next node is then evaluated. The path followed is called the evaluation path.

The DDAG method basically eliminates one class out from a list. Initially the list contains all classes. Each node evaluates the first class against the last class in the list. For example, the root node evaluates class 1 against class $M$. If the evaluation results in one class out of two classes, the other is eliminated from the list. The process then tests the first and the last class in the new list. It is terminated when only one class remains in the list. The class label associated with the input data will be the class label of the node in the final layer of the evaluation path or the class remained in the list. Although the number of binary classifiers still equals the pairwise classification method, the inputs are evaluated $M-1$ times instead of $0.5^*M(M-1)$ times as is the case with pairwise classification.
4.2.4 Multiclass Objective Function

Instead of creating many binary classifiers to determine the class labels, this method attempts to directly solve a multiclass problem. This is achieved by modifying the binary class objective function and adding a constraint to it for every class. However, in this method, the optimization algorithm has to consider all the support vectors at the same time. Therefore, it may be able to handle massive data sets but the memory requirement and thus, the computational time may be very high. A user should consider the accuracy requirement, the computational time, the resources available and the nature of the problem. For example, the multiclass objective function approach may not be suitable for a problem that contains a large number of training samples and classes due to the requirement of large memory and extremely long computational time.

4.2.5 Error-Correcting Output Code based approach

The concept of Error-Correcting Output Coding (ECOC) based multi-class method is to apply binary (two-class) classifiers to solve the multi-class classification problems. This approach works by converting \( M \) class classification problem into a large number \( L \) of 2-class classification problems. ECOC assigns a unique code word to a class instead of assigning each class a label. A \((L, M, d)\) error correcting code is \( L \) bit long, having \( C \) unique code words with a Hamming distance of \( d \). The hamming distance between two code words is the number of bit positions in which both differs. In a classification problem \( M \) is the number of classes and \( L \) is a number decided by the method used to generate error-correcting codes.
4.3 MULTICLASS CLASSIFICATION BY SVM

Originally, SVMs were developed to perform binary classification. 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. A support vector machine finds an optimal separating hyperplane between members and non-members of a given class in an abstract space. In addition to counteracting over fitting, the SVM's use of the maximum margin hyperplane leads to a straightforward learning algorithm that can be reduced to a convex optimization problem. In order to train the system, the SVM must find the unique minimum of a convex function. Furthermore, the support vectors identified by the SVM effectively define the boundary of the training set of genes. This ability to focus on the few informative genes out of the vast landscape of uninformative genes is fundamental to making scientific insight. One significant benefit offered by SVMs is scalability. The number of support vectors selected by the SVM learning algorithm is usually small, even for very large training sets, and the resulting SVM is consequently an efficient classifier. Thus, classifying a new gene requires comparisons with only approximately 13.1% of the training set.

Scalability is essential because the amount of available gene expression data will soon increase dramatically. When hundreds, and soon thousands, of mRNA expressions measurements under different conditions become available for each gene, each measurement will still, by itself, give only partial and inconclusive information about any given functional classification of the gene. However, all these different mRNA measurements taken together may often provide enough information to classify the gene with very high confidence. It uses only DNA microarray expression data, but similar SVMs could be constructed using other gene features, such as the presence of transcription factor binding sites in the promoter region or sequence features of the translated protein suggests Alireza & Bita (2009).
A simple and effective approach with a given m classes trains m classifiers, one for each class (where classifier j learns to return a positive value for class j and a negative value for the rest). A test tuple is assigned the class corresponding to the largest positive distance. Finding most predictive features for statistical classification is a challenging problem and has important applications. Support Vector Machines (SVMs), for example, have been found successful with a recursive procedure in selecting most important genes for cancer prediction. This is by examining multiple classifiers SVM with feature selection in recursive and non-recursive settings, on a DNA microarray dataset (AML or ALL) and a text categorization benchmark. Jing et al (2006) portrays saying more specifically choose to examine the linear version of the classification methods, since linear classifiers are relatively simple, easy to interpret, and can be enriched through the use of kernel functions for solving non-linear problems.

4.3.1 Applications using Multiclass SVM

4.3.1.1 Molecular biology research

It is not possible to research on a large number of genes using traditional methods. DNA Microarray is one such technology which enables the researchers to investigate and address issues which were once thought to be non traceable. One can analyze the expression of many genes in a single reaction quickly and in an efficient manner. A typical microarray experiment involves the hybridization of an mRNA molecule to the DNA template from which it is originated. Many DNA samples are used to construct an array. The amount of mRNA bound to each site on the array indicates the expression level of the various genes. This number may run in thousands. All the data is collected and a profile is generated for gene expression in the cell. Thousands of spotted samples known as probes (with known identity) are immobilized on a solid support (a microscope glass slides or silicon chips or nylon membrane). The spots can be DNA, cDNA, or oligonucleotides.
These are used to determine complementary binding of the unknown sequences thus allowing parallel analysis for gene expression and gene discovery. SVM and validation of cancer tissue samples using microarray gene data is analyzed with ovarian dataset. With classification and validation of ovarian cancer tissues, normal ovarian tissues and other normal tissues like data sets are adopted in Terrence et al (2000). Upon correction of this mistake and the removal of an outlier, perfect classification of tissues is achieved, but not with high confidence. Identify and analyze a subset of genes from the ovarian dataset whose expression is highly differentiated between the types of tissues.

4.3.1.2 Face recognition

In the face recognition, more than two face classes are used. Each SVM is a binary-class classifier, so, in order to classify more than two classes, multi-class SVMs are needed. There are two basic schemes to classify $K$ face classes when using multi-class SVMs. The first one is "one vs. all" scheme that composes $K$ binary-class SVMs to classify one class from all classes, and the second one is "one vs. one" scheme that composes all pairs of classes for binary-class SVMs. In the one-vs.-one scheme, with $K$ training data, $K (K-1)/2$ SVMs are trained to separate two face classes. First, the one-vs.-one scheme is simply constructed. Although one vs. one scheme has a fault which tends to build many binary-class SVMs, this scheme is excellent at training cost compared to the one-vs.-all scheme.

Also, the one-vs.-all scheme can lead to an ambiguous classification process. The one-vs.-one scheme can use both bottom-up and top-down tree structures and the bottom-up structure demonstrate excellent performance in 3D object recognition and face recognition. Most commonly, an individual class consists of a fixed length vector (also known as its chromosome), and the individual’s fitness is computed directly from this vector. If any individual is found with a fitness exceeding a predetermined threshold and that individual is
returned as the solution to the problem. The run is also terminated if the maximum number of generations have been read, in which case the fittest individual is found and returned. If the run is not terminated, a new generation is read. As the database of facial images gets bigger, both the numbers of training for SVMs and comparisons increase, and the classification time greatly lengthens.

4.3.1.3 Remote sensing

The applications involved multiclass classification, especially in remote sensing land cover. It compares the performance of six multi-class approaches to solve classification problem with remote sensing data in term of classification accuracy and computational cost. One vs. one, one vs. rest, Directed Acyclic Graph (DAG), and Error Corrected Output Coding (ECOC) based multiclass approaches creates many binary classifiers and combines their results to determine the class label of a test pixel. This classification technique is to separate the classes with a surface that maximize the margin between them, using boundary pixels to create the decision surface. This problem of maximizing the margin can be solved using standard Quadratic Programming (QP) optimization techniques.

If the two classes are not linearly separable, the SVM tries to find the hyperplane that maximizes the margin while, at the same time, minimizing a quantity proportional to the number of misclassification errors. The trade-off between margin and misclassification error is controlled by a user-defined constant. SVM can also be extended to handle non-linear decision surfaces. A second approach included in (Mahesh 2008) is to combine several classifiers (‘one against one’) perform pair-wise comparisons between all n classes. Thus, all possible two class classifiers are evaluated from the training set of n classes, each classifier being trained on only two out of n classes, giving a total of n (n-1)/2 classifiers. For M classes, there will be binary classifiers. The output from each classifier in the form of a class label is obtained. One against one and DAG approach provide a comparable accuracy and requires almost same computational
resources. The training time taken by one against one and DAG techniques is less than that with the one against the rest strategy.

4.3.2 Methods and Analysis

It is evident that applications of nonlinear dimension reduction techniques could have a promising perspective in microarray data analysis. PCA has been mainly used to analyze spatial data and also a useful tool for time series data. Since PCA retains maximum variance, it is expected to provide features that are robust to small noise. In applications of support vector machines to cancer classification with microarray data the support vector machine (SVM) for cancer classification with microarray data are used. Dimensionality reduction methods, such as principal components analysis (PCA), class-separability measure, Fisher ratio, and $t$-test, are used for gene selection. A voting scheme is then employed to do multi-group classification by $k$ ($k - 1$) binary SVMs. The classification accuracy with much fewer features were obtained. Microarrays also known as gene chips or DNA chips provide a convenient way of obtaining gene expression levels for a large number of genes simultaneously. Each spot on a microarray chip contains the clone of a gene from a tissue sample. It is challenging to use gene expression data for cancer classification because of the following two special aspects of gene expression data.

A powerful classifier, i.e., the support vector machine (SVM), and four effective feature reduction methods, i.e., principal components analysis (PCA), class-separability measure, Fisher ratio, and $t$-test, to the problem of cancer classification based on gene expression data. The entire data set includes the expression data of 2308 genes. There are totally 63 training samples and 25 testing samples. There are 42 samples derived from diffuse large B-cell lymphoma (DLBCL), 9 samples from follicular lymphoma (FL), and 11 samples from chronic lymphocytic lymphoma (CLL). In this data set, a small part of data is missing. A $k$-nearest neighbor algorithm was applied to fill those missing values Feng & Lipo (2005). The samples in this data set belong to two types of leukemia, i.e., the acute
myeloid leukemia (AML) and the acute lymphoblastic leukemia (ALL). Among these samples, 38 of them are used for training and the other 34 independent samples are for testing.

Xiyi & Fang (2009) have proposed a technique for cancer diagnosis by gene expression data. The proposed technique has represented each testing sample as a linear combination of all the training samples. The coefficient vector has been obtained by $l_1$-regularized least square. Classification has been performed by defined discriminating functions using the coefficient vector for the meta-samples extracted from each category. $l_1$-Norm minimization has produced a sparse solution, and they called this approach as sparse representation. Experimental results have proved that the sparse representation approach was very efficient in tumor classification.

Mark et al (2008) have proposed a technique that considered a larger subset of CAR-related (conjunctive association rules) and Boolean Association Rules (BARs). To deal with the computational complexities included with pre-classification CAR mining, those rules have been efficiently captured in a Boolean Structure Table (BST), which has been then utilized to create a BST classifier called BSTC. In comparison to the present foremost CAR classifier, RCBT, on many benchmark microarray datasets have confirmed that the BSTC is competitive with RCBT’s accuracy while minimizing the exponential costs acquired by CAR mining. Thus, BSTC extended the generalized CAR-based techniques to larger datasets. Besides, contrasting from other association rule-based classifiers, BSTC easily generalized to multi-class gene expression datasets. BSTC’s worst case per-query classification time was poorer than CAR-based techniques ($O(|S|2 \cdot |G|)$ versus $O(|S| \cdot |G|)$).

If there are more than two classes in the data set, binary SVMs are not sufficient to solve the whole problem. To solve multi-class classification problems, we should divide a whole problem into a number of binary classification problems. Usually, there are two approaches. One is the “one against all” scheme and the other is the “one against one” scheme. For the “one against all” scheme, if there
are $N$ classes in the entire data set, then $N$ independent binary classifiers are built. For the “one against one” scheme, there must be one (and only one) classifier taking charge of the classification between any two classes. Therefore, for a data set with $K$ classes, $K(K-1)/2$ binary classifiers are used. To get the ultimate result, a voting scheme is used.

For every input vector, all the classifiers give their votes so there will be $K(K-1)/2$ votes, when all the classification (voting) finished, the vector is designated to the class getting the highest number of votes. The input data set genes one by one to the SVM classifier according to their ranks. Then, it is trained with the SVM classifier with the training data and tested the SVM classifier with the testing data implemented by Feng & Lipo (2005). After that, the whole process with top 2 genes, and then top 3 genes, and so on are repeated. In this data set, we used SVMs with RBF kernels. $C$ and $\gamma$ were set as 80 and 0.005, respectively. This classifier obtained 100% training accuracy and 100% testing accuracy using the top 7 genes. Linear SVMs and SVMs with polynomial kernel function to the data set were applied. The linear SVMs and the SVMs with the polynomial kernel function obtained 100% accuracy with 7 and 6 genes, respectively.

4.4 DATASET DESCRIPTION

In this section, the assessment based on different microarray data sets is presented. These datasets are of high density, it is preprocessed with the gene expression data using the standard procedure including background correction, transformation and normalization. The experimentation of these datasets with SVMs as well as with other learning techniques is a large scale process. Datasets are used in the appropriate way based on different methods and algorithms. Discriminant analysis algorithms are originally designed for continuous valued data sets. However, with simple preprocessing, it can be used on any type of data sets too. In this section the details of the data sets are described.
4.4.1 Ovarian Dataset

Ovarian cancer is a cancerous growth arising from the ovary. Most (more than 90%) ovarian cancers are classified as "epithelial" and are believed to arise from the surface (epithelium) of the ovary. The majority of ovarian cancers arise from the epithelium (outer lining) of the ovary. Expression experiments are performed using 97802 DNA clones, each of which may or may not correspond to human genes, for 31 tissue samples. From these samples 1000 features are created. Experiments using similar diagonal factors to those are performed using these smaller feature sets. Applications display the most significant results from these experiments. The best classification is done using the top 50 features with a diagonal factor of 2 or 5. Though the smaller datasets achieve slightly better scores compared to using all features and this improvement to be significant. An analysis of the misclassified examples reveals that normal ovarian tissue sample. In addition, the margin of misclassification, calculated is relatively large meaning the SVM strongly believes it to be cancerous.

4.4.2 The Lymphoma Dataset or AML/ALL Dataset

The leukemia data samples belong to two types of leukemia, i.e., the acute myeloid leukemia (AML) and the acute lymphoblastic leukemia (ALL).

There are 4 main types of leukemia:

- Acute myeloid (or myelogenous) leukemia (AML)
- Chronic myeloid (or myelogenous) leukemia (CML)
- Acute lymphocytic (or lymphoblastic) leukemia (ALL)
- Chronic lymphocytic leukemia (CLL)

Bone marrow or peripheral blood samples are taken from 72 patients with either acute myeloid leukemia (AML) or acute lymphoblastic leukemia (ALL). Acute lymphoblastic leukemia (ALL) is the most common type of leukemia in young children. This disease also affects adults, especially those ages 65 and older. Standard treatments involve chemotherapy and radiotherapy. Acute lymphocytic
leukemia (ALL), also called acute lymphoblastic leukemia, is a cancer that starts from white blood cells called lymphocytes in the bone marrow (the soft inner part of the bones, where new blood cells are made). The term “acute” means that the leukemia can progress quickly, and if not treated, would probably be fatal within a few months. Lymphocytic means it develops from early (immature) forms of lymphocytes, a type of white blood cell. The survival rates vary by age: 85% in children and 50% in adults. “Myeloid” refers to the type of cell the leukemia starts from. AML is a cancer that starts in cells that would normally develop into different types of blood cells.

Most cases of AML develop from cells that would turn into white blood cells (other than lymphocytes), but some cases of AML develop in other types of blood-forming cells. AML starts in the bone marrow (the soft inner part of the bones, where new blood cells are made), but in most cases it quickly moves into the blood. It can sometimes spread to other parts of the body including the lymph nodes, liver, spleen, central nervous system (brain and spinal cord), and testicles.

Acute myelogenous leukemia (AML) occurs more commonly in adults than in children, and more commonly in men than women. AML is treated with chemotherapy. Subtypes of AML include acute promyelocytic leukemia, acute myeloblastic leukemia, and acute megakaryoblastic leukemia.

The entire leukemia data set contains the expression data of 7129 genes. Ordinarily, raw gene expression data should be normalized to reduce the systemic bias introduced during experiments was demonstrated by Nevine & Michael (2007). The acute leukemia data contain 72 bone marrow samples on 7129 genes from patients with either acute lymphoblastic leukemia (ALL) or acute myeloid leukemia (AML). The original data consist of a training set of 38 samples with 27. However, for the leukemia data set, normalized data are not available. Thereafter normalization is necessary. In this data set, a small part of data is missing. A k-nearest neighbor algorithm is applied to fill those missing values.
**Chronic leukemia:** In chronic leukemia, the cells can mature partly but not completely. These cells may look fairly normal, but they generally do not fight infection as well as normal white blood cells do. They also live longer, build up, and crowd out normal cells. Chronic leukemias tend to progress over a longer period of time, and most people can live for many years. But chronic leukemias are generally harder to cure than acute leukemias.

### 4.4.3 Colon Tumor Dataset

The colon is the longest part of the large intestine and the lowest part of the digestive system. Colon cancer forms when this uncontrolled cell growth initiates with cells in the large intestine. Most colon cancers originate from small, noncancerous (benign) tumors called adenomatous polyps that form on the inner walls of the large intestine. Using oligonucleotide arrays, expression levels for 40 tumor and 22 normal colon tissues are measured for 6500 human genes. Of these genes, the 2000 with the highest minimal intensity across the tissues are selected for classification purposes. Each score represents a gene intensity derived in a process. The data is not processed further before performing classification. The experiments use a clustering method to create clusters of tissues. In their experiments, one cluster consists of 35 tumor and three normal tissues, and the other 19 normal and five tumor tissues. The colon data set has 62 tissue samples on 2000 genes.

### 4.4.4 IRIS Dataset

A classification data set based on characteristics of a plant species (length and thickness of its petal and sepal) divided into three distinct classes (*Iris Setosa, Iris Versicolor* and *Iris Virginica*). The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant. One class is linearly separable from the other 2; the latter are not. The predicted attribute is class of iris plant. In total there are 150 records each with 4 attributes and 50 records belongs to each
class mentions Satchidananda & Cho (2008). The database is divided into two sets each containing 75 total records and 25 from each class.

4.4.5 CLOUD Dataset

This dataset is derived from AVHRR images (Advanced Very High Resolution Radiometer). It consists of 2048 instances in 10 dimension distributed into two classes.

4.4.6 HABERMAN Dataset

This dataset is based on the survival of patient’s undergone surgery for breast cancer consists of 306 instances in 3 dimensions distributed into 2 classes.

4.4.7 HEART Dataset

This dataset derived from the diagnoses of people with heart problem consists of 270 instances in 13 dimension distributed into 2 classes.

4.4.8 PIMA Dataset

The Pima Indian diabetes dataset based on clinical tests consists of 768 instances in 8 dimensions distributed into 2 classes. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

4.4.9 The SRBCT Dataset

The small round blue cell tumors (SRBCTs) are 4 different childhood tumors named so because of their similar appearance on routine histology, which makes correct clinical diagnosis extremely challenging. However, accurate diagnosis is essential because the treatment options, responses to therapy and prognoses vary widely depending on the diagnosis. The SRBCT data set includes
the expression data of 2308 genes. There are totally 63 training samples and 25 testing samples, five of the testing samples being not SRBCTs.

4.5 MULTICLASS SVM ENSEMBLE

Classifying genes is one of the tough tasks to perform with the aid of microarray because of its high dimensional data. After that $\eta$ SVM classifiers are generated for $\eta$ classes. The technique consists of two phases which are dimensionality diminution and the classification based on the multi SVM.

4.5.1 Dimensionality Diminution through Principal Component Analysis (PCA)

![Diagram of multiclass SVM ensemble classifier for microarray gene data classification]

Figure 4.17 Training process of multiclass SVM ensemble classifier for microarray gene data classification
Prior SVM classification, PCA is utilized to reduce the dimensionality of the microarray gene data. Let $M_{xy}; 0 \leq x \leq n_g, 0 \leq y \leq n_s$ here $n_g$ indicates the number of genes in the sample in which it has been taken from and $n_s$ indicates the number of samples in which has been taken for the process. These microarray gene data is of higher dimension and hence it must be reduced in order to do that the PCA mechanism is utilized. In this dimensionality diminution, data which is in high dimension is converted to low dimension. After that the resulted data is given for the classification process, multi SVM is utilized for this process. These SVMs are trained well with the aid of reduced dimension microarray data and here the more than a single class is trained and identified.

The dimensionality diminution is the process of reducing the large dimensional data, in order to make comfort to the classification process. PCA is one of the dimensionality reduction techniques that are utilized. PCA is a great tool for the data analysis process and also it can be utilized for the dimensionality reduction without any loss of information. The following steps are the dimensionality reduction steps of our microarray gene data.

Step 1: Compute the mean of the microarray gene data

$$
\mu = \frac{1}{N_s \times N_g} \sum_{i=1}^{N_s} \sum_{j=1}^{N_g} M_{ij}
$$

(4.1)

Step 2: Subtract the mean from each gene data to find the mean deviation

$$
\delta = M_{xy} - \mu
$$

(4.2)

Step 3: Compute the covariance matrix for the data $\delta$

$$
Cov = \frac{\delta \times \delta^T}{y-1}.
$$

(4.3)

Step 4: Compute the Eigen values and Eigen vector and determine

$$
\lambda = \mu \times E^T
$$

(4.4)
Step 5: After computing the Eigen values and vector then the embedding process is as follows in

\[
\hat{M} = \frac{\hat{A}}{E} \ast \mu
\]  

(4.5)

After obtaining the PCA transformed values then the inverse PCA transformation is also applied in order to diminution process on the microarray gene data. Once the dimensionality reduced data are applied with the SVM in order to classify them. The multi class SVM ensemble is utilized for the classification process. The following session of this section details the classification process for multi class SVM ensemble.

### 4.5.2 Multiclass SVM Ensemble for Microarray Gene Data Classification

The SVM is trained with the aid of the dimensionality reduced data in order to classify them. Here we utilize SVM ensemble for classifying multi class microarray gene data. For this SVM ensemble \( \eta \) SVMs are generated for the \( \eta \) classes as discussed earlier. SVMs are related to the widespread classifiers category. SVMs are also regarded as a special case of Tikhonov regularization. The unique property of the SVM is it has diminished the empirical classification errors and boosts up the geometric margin simultaneously.

In this, \( \eta \) classifiers for \( \eta \) classes and which may be utilized to identify the class of a gene more strongly than the single SVM is presented. Here 3 class of gene data are utilized where they are ALL, AML and Lymphoma datasets. The SVM ensemble is the process of making a strong decision about the class of an input test gene. Here three SVMs for three different gene datasets are generated. The generated SVMs are trained with the aid of the dimension diminished gene data.
Figure 4.18 Multi class SVM ensembles for microarray gene data classification

The following equations are utilized to reduce the error and hence it trains the multi class SVM effectively.

\[ Err = \arg \min_{\rho} \sum_{n=0}^{N-1} \omega_n + 0.5\psi^T \psi \]  
\[ (4.6) \]
\[ c_y (\psi^T K (\mathbf{M}) + v) \geq 1 - \omega_n \]  \hspace{1cm} (4.7)

\[ \omega_n \geq 0 \]  \hspace{1cm} (4.8)

In the above set of equations, \( \rho \) is the penalty constant, \( \omega \) is the attribute utilized to handle the data and \( c_y \) is the class of the \( y^{th} \) dataset, \( \psi \) is the matrix of coefficients. \( K \) is the kernel that renovates the input data to the feature space and \( v \) is the constant. By diminishing the error function, the SVMs learns the training microarray gene dataset Theses multiclass SVM ensemble is trained well and hence they are ready to categorize any gene data belongs to this similar fashion of genes

**Classification Process:** In this classification process, a test input gene data is utilized in order to classify and for this process, initially, the input gene dataset is applied with the PCA for diminishing its dimension. Once the dimension is reduced it may given to the SVM ensemble. The process of the SVM ensemble is to identify the class of the gene. Here the SVM ensemble may indicates a strong decision of the class of a gene data.

**Input:** Test input gene data

**Output:** class of test gene data

**Steps**

**Step1:** Test the input test gene in SVM1

Obtained class is C1

**Step2:** Test the input test gene in SVM2

Obtained class is C2

**Step3:** Test the input test gene in SVM3

Obtained class is C3

**Step 4** Identify the frequency of classes

**Step 5:** Highly frequented class
The above steps detail the process of classification process. Hence the highly frequented class is observed as the class of test input gene is the class of the gene. Thus the multi class SVM ensemble is utilized to identify the class of the gene. The technique is used for classifying microarray gene data and this is implemented in the working platform of MATLAB (Version 7.11). The technique is assessed with the aid of gene samples of Lymphoma, Acute Myeloid leukemia (AML) and Acute Lymphoblastic Leukemia (ALL). The aforementioned gene data is reduced with the aid of PCA technique and then it has been applied with the training process of 3 SVMs individually. In the testing process, the test gene data is tested in each of the SVMs and then it has given the better results. The gene data for dimension is as follows:

Table 4.2 Microarray gene data dimension utilized for the evaluation process

<table>
<thead>
<tr>
<th>Type of Gene Data</th>
<th>Number of Samples</th>
<th>Number of Genes</th>
<th>Dimensionality Reduced Data with the aid of PCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL</td>
<td>27</td>
<td>7030</td>
<td>27 X 38</td>
</tr>
<tr>
<td>AML</td>
<td>11</td>
<td>7030</td>
<td>11 X 38</td>
</tr>
<tr>
<td>Lymphoma</td>
<td>58</td>
<td>4000</td>
<td>58 X 38</td>
</tr>
</tbody>
</table>

The aforementioned table shows that the input sample data, their dimension and the reduced dimension is mentioned. Here sample of microarray gene dataset of 3 classes has been utilized for the training process. In this SVM ensemble, the ALL-AML, AML-Lymphoma and Lymphoma-ALL are trained individually in three SVMs. Once the training process completed, in the testing process, the input test gene is given and they are analyzed in each SVMs and the majority class is
identified as the class of the input microarray gene data. The following are the formulas for the obtained results parameters, accuracy, sensitivity, specificity and the error rate.

\[
\text{Accuracy} = \frac{\text{No. of samples classified exactly}}{\text{Total number of samples subjected to classification}} \tag{4.9}
\]

\[
\text{Sensitivity} = \frac{\text{No. of true possitives}}{\text{No of true positives + No of False negatives}} \tag{4.10}
\]

\[
\text{Specificity} = \frac{\text{No. of true negatives}}{\text{No of true negatives + No of False positives}} \tag{4.11}
\]

\[
\text{Error rate} = 1 - \frac{\text{No. of samples classified exactly}}{\text{Total number of samples subjected to classification}} \tag{4.12}
\]

Table 4.3 Performance comparison for the classified data with the aid of dimensionality reduced data (PLS)

<table>
<thead>
<tr>
<th>Type of Gene Data</th>
<th>Specificity</th>
<th>Sensitivity</th>
<th>Accuracy</th>
<th>Error Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL</td>
<td>0.2222</td>
<td>0</td>
<td>0.0862</td>
<td>0.9138</td>
</tr>
<tr>
<td>AML</td>
<td>0</td>
<td>0.0862</td>
<td>0.2222</td>
<td>0.7778</td>
</tr>
<tr>
<td>LYMPHOMA</td>
<td>0.1579</td>
<td>0.0725</td>
<td>0.1294</td>
<td>0.8706</td>
</tr>
</tbody>
</table>
Figure 4.19 Performance comparison for the classified data with the aid of dimensionality reduced data (PLS)

Table 4.4 Performance comparison for the classified data with the aid of dimensionality reduced data (PCA)

<table>
<thead>
<tr>
<th>Type of Gene Data</th>
<th>Specificity</th>
<th>Sensitivity</th>
<th>Accuracy</th>
<th>Error Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL</td>
<td>0.9259</td>
<td>0.5455</td>
<td>0.9310</td>
<td>0.0690</td>
</tr>
<tr>
<td>AML</td>
<td>0.5455</td>
<td>0.9310</td>
<td>0.9259</td>
<td>0.0741</td>
</tr>
<tr>
<td>Lymphoma</td>
<td>0.8158</td>
<td>0.8696</td>
<td>0.9294</td>
<td>0.0706</td>
</tr>
</tbody>
</table>

Figure 4.20 Specificity, Sensitivity and Accuracy of the Input samples ALL, AML and Lymphoma with the aid of proposed work
Table 4.3 details the sensitivity, specificity, accuracy and error rate for the multi class SVM ensemble with the PLS dimensionality reduction algorithm. Figure 4.19 depicts the work with the performance measures sensitivity, specificity and the error rate for the ALL, AML and the Lymphoma dataset. Table 4.4 depicts the performance measure for the proposed work and figure 4.20 depicts the performance measures for this work with 3 different classes ALL, AML and Lymphoma. Thus the results state that the present work classified the input microarray gene data in an effective manner.

**Robustness analysis:** The large number of samples as shown in Table 4.2 here allows assessing the robustness of the differentially regulated genes between the sample groups. The approach PCA and PLS performed a sampling approach and both show the varying robustness of differentially regulated genes for different group comparisons. The test for robustness was implemented to maintain continued performance against intrinsic perturbations and uncertainty of living systems and has been reviewed extensively. Considering the cutoff value PLS covers> 82% of all genes and PCA covers > 93% in the strict analysis. The results of this analysis show that the values of PCA are generally very robust.

### 4.6 SUMMARY

An effective microarray gene data classification technique with the aid of PCA and SVM is presented in this chapter. The dimensionality of the microarray data has been reduced with the aid of PCA mechanism. After that the SVM ensemble is utilized for the classification process in order to classify more than a single class. In the testing process any of these genes given and they have identified in which the class they belongs to. The experiment described has obtained better results and they are compared with another dimensionality reduction algorithm i.e. PLS. The forth coming chapter will discuss in detail about the efficient retrieval techniques in classification and clustering.