DOKUZ EYLÜL UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

CLASSIFICATION OF WISCONSIN BREAST CANCER DATABASE

by
Cihan GÜNEŞER

September, 2009
İZMİR

CLASSIFICATION OF WISCONSIN BREAST CANCER DATABASE

A Thesis Submitted to the
Graduate School of Natural and Applied Sciences of Dokuz Eylül University
In Partial Fulfillment of the Requirements for the Degree of Master of Science
in Electrical and Electronics Engineering

by
Cihan GÜNEŞER

September, 2009
İZMİR

M. Sc THESIS EXAMINATION RESULT FORM

We have read the thesis entitled **"CLASSIFICATION OF WISCONSIN BREAST CANCER DATABASE"** completed by **CİHAN GÜNEŞER** under supervision of **Asst. Prof. Dr. METEHAN MAKİNACI** and we certify that in our opinion it is fully adequate, in scope and in quality, as a thesis fort he degree of Master of Science

---

Asst. Prof. Dr. Metehan MAKİNACI

Supervisor

……

…..

Prof. Dr. Cahit HELVACI

Director

Graduate School of Natural and Applied Sciences

ACKNOWLEDGMENTS

I would like to state that this study has been developed faster, under counseling of my advisor Asst. Prof. Dr. Metehan MAKİNACI. Especially his ideas for choosing the right materials and comparison of methods were quite useful. I also would like to thank my family and managers for their support.

Cihan GÜNEŞER

CLASSIFICATION OF WISCONSIN BREAST CANCER DATABASE

ABSTRACT

Statistical data analysis includes developing methods for classification of various databases. These databases may have data about lots of sectors for example financial, industrial, food, biomedical or etc. The main aim is to get a result by making a classification for a product, patient or something.

In this study, we used classification methods for biomedical analysis. Our sample database has breast cancer data which is one of the most cause of cancer, because detection of this cancer is very important. Database has 9 attributes which is used for classification. These attributes are numerical numbers. Explanations of attributes are given in detail in following chapters. After having numerical numbers via FNA procedure, class labels can be given both by classical examinations and by classification algorithms. Our database includes a class column which real diagnosis exists. This study aims to consider classification algorithms results carefully.

Different methods and algorithms have been used; classification accuracies have been given depending on real values. Results are compared and some ideas can arise for using software programs to classify sickness instances. Computer supported diagnosis can be used more commonly.

Keywords: Biomedical data classification, KNN rule, linear discriminant, neural network, support vector machine, breast cancer.

WISCONSIN MEME KANSERİ VERİTABANI SINIFLANDIRMASI

ÖZ

İstatiksel veri analizi, farklı veritabanlarında sınıflandırma yapabilmek için metot geliştirmeyi içerir. Bu veritabanları farklı sektörlere ait olabilir. Örneğin finans, endüstri, gıda, biyomedikal vb. Asıl amaç bir hastayı, bir ürünü veya herhangi bir şeyi sınıflandırarak bir sonuç elde etmektir.

Bu çalışmada , biyomedikal analiz için sınıflandırma metotlarını kullandık. Örnek veritabanımız, kanserin en büyük sebeplerinden biri olan meme kanseri verilerini içermektedir. Çünkü meme kanserinin teşhisi büyük önem taşımaktadır. Veritabanımızda, sınıflandırma için kullanılacak 9 öznitelik bulunmaktadır. Bu öznitelikler sayılardan oluşmaktadır. İlerleyen bölümlerde bu özniteliklerin detayları verilecektir. FNA işlemi sonucunda sayısal veriler alındıktan sonra, sınıflandırma etiketleri klasik yöntemlerle veya sınıflandırma algoritmaları ile verilebilir. Veritabanımızda gerçek tanı değerleri bulunan bir sütun bulunmaktadır. Bu çalışma, sınıflandırma algoritmalarının önemini vurgulamaktadır.

Çalışmada farklı algoritmalar ve metotlar kullanılmış, gerçek sonuçlar baz alınarak hesaplanan doğruluk oranları verilmiştir. Elde edilen sonuçlar karşılaştırılmıştır. Böylece, bilgisayar programlarının, hastalık tanılarında kullanılması için yeni fikirler oluşabilir. Bilgisayar destekli tanı işlemi daha yaygın şekilde kullanılabilir.

Anahtar Kelimeler: Biyomedikal veri analizi, KNN kuralı, linear discriminant, neural network, support vector machine, meme kanseri.

CONTENTS

Page

CHAPTER ONE
INTRODUCTION

1.1 Introduction

Classification algorithms are for finding valuable information in big databases. In health care industry, these databases are saved in computers, so they can be used for improving methods of classification.

Biomedical data are the clues for diagnosis of our sickness. They can be diagnosed by some different methods. Some of them are biopsy, fine needle aspirate, and mammography. Most of them are based on taking samples of blood or piece of our body. All these methods have different accuracy. Biopsy has the best accuracy between them. FNA and mammography comes orderly after biopsy. All three of these methods may be used for diagnosing breast cancer. Before getting these numerical attributes; there is Fine Needle Aspiration (FNA) process. FNA is used to investigate lumps or masses of cells. For known tumors, this biopsy is usually performed to assess the effect of treatment or to obtain tissue for special studies.

These biomedical data which are determined by these methods may be variables including numbers or labels of classes. Some of the attributes are more important than others. And also they have bigger role for having a result about sickness. For this reason, in some classification methods, they have been taken in to consideration more than others attributes.

In this thesis, our sample database is a breast cancer database, which includes attributes, class label and sample code number. Details of this database will be explained in other chapters. We used four different methods for analyzing our database: Linear Discriminant, Support Vector Machine, K Nearest Neighborhood, and Neural Network. These methods have different ways of analyzing and classifying data. They have different priorities for having a result.

As we are having different values for every single classification, we could have misleading results. This is because of structure of samples and numerical values of attributes. For this reason, we calculated performances for 50 times for every four algorithms. Train set was different in each calculation (cross-validation). Average performance values were obtained from these results. These results are the nearest values to real accuracy.

In Chapter 2, details are given on methods. Mathematical models and formulas are given. In Chapter 3, dataset is explained with details, attribute explanations are given. Classification results are explained with confusion matrices. Our four accuracy groups are compared with each other. These overall accuracy values are: 96.6% for linear discrimination, 97.3% for support vector machine, 97.2% for k nearest neighborhood rule and 96.06% for neural network. Some other studies are examined and results are given in the same chapter. These studies and results are: 99.20% made by Mihir Sewak with SVM for Wisconsin Breast Cancer Database; 97.90% made by I. Anagnostopoilosi by using advanced neural network techniques, as well as Z. Yang had 98.50%. Y.Li, T. Mu, W. Wolberg made a classification application for Wisconsin Breast Cancer Database. Accuracy results are 95.50%, 98.40% and 97.50% respectively.

In Chapter 4, conclusions are given, and some ideas are shared for future work.

CHAPTER TWO
METHODS

In this chapter, methods are explained in an order. Details of linear discriminant functions, support vector machine, KNN rule and neural network are given. Reader can find the details of methods in reference books (Duda, Hart & Stork 2001), (Hsu & Chang 2008), (Yang, Lu, Yu & Harrison 2000)

2.1 Linear Discriminant

Linear discriminant analysis (LDA) and Fisher's linear discriminant are widely used methods in statistics. They are used to find linear combination of features which best separate two or more classes of objects or events. The results are used as a linear classifier for classification.

In LDA, the dependent variable is a categorical variable, like class label. In analysis of variance and regression analysis, we attempt to express one dependent variable as a linear combination of features but the dependent variable is a numerical quantity. Discriminant analysis is also different from factor analysis in that it is not an interdependence technique: a distinction between independent variables and dependent variables (also called criterion variables) must be made. LDA works with measurements which are continuous quantities. If the variables are categorical, the equivalent technique is discriminant correspondence analysis.

2.1.1 LDA for Two Classes

Consider a set of observations x (also called features, attributes, variables or measurements) for each sample of an object or event with known class y. The set of samples is called the training set. The main subject is to find a good predictor for the class y of any sample.

LDA approaches the problem by assuming that the conditional probability density functions are both normally distributed. Under this assumption, the Bayes optimal solution is to predict points as being from the second class if the likelihood ratio is below some threshold T.

The resulting classifier is quadratic discriminant analysis. LDA makes the simplifying that the class covariance is identical and that have full rank. In this case, several terms cancel and the above decision criterion becomes a threshold on the dot product. This means that the probability of an input *x* being in a class *y* is purely a function of this linear combination of the known observations.

A discriminant function that is a linear combination of the combination of the components of *x* can be written as

$$g(X) = W^t X + w_0 \qquad \text{Eq. 2.1}$$

where **w** is the weight vector and w_0 the bias or threshold weight. A two-category linear classifier implements the following decision rule: Decide w_1 if $g(x)>0$ and w_2 *if* $g(x)>0$. Thus, *x* is assigned to w_1 if inner product $\mathbf{w}^t x$ exceeds the threshold $-w_0$ and w_2 otherwise. If $g(x) = 0$, *x* can ordinarily be assigned to either class. Figure 2.1 shows a typical implementation.

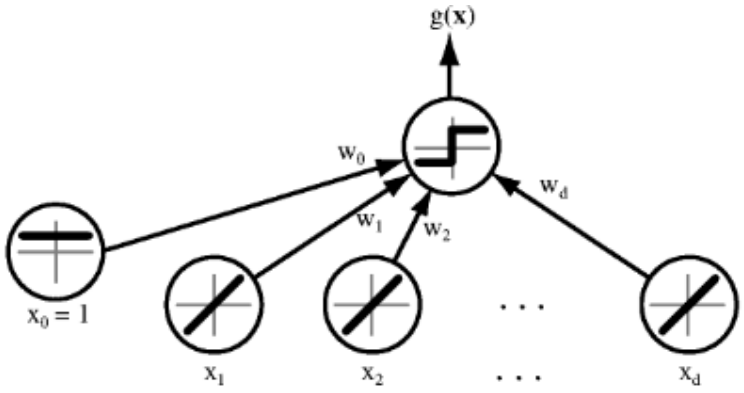

Figure2.1 A simple linear classifier having *d* input units, each corresponding to the values of the components of an input vector.

To summarize, a linear discriminant function divides the feature space by a hyper plane decision surface. The orientation of the surface is determined by the normal vector **w,** and the location of the surface is determined by the bias w_0. The discriminant function $g(x)$ is the proportional to the signed distance from x to hyper plane, with $g(x)>0$ when x is on the positive side, and $g(x)<0$ when x is on the negative side. (Duda, Hart & Stork 2001)

2.1.2 Fisher's Linear Discriminant

Fisher's linear discriminant is a classification method that projects high-dimensional data onto a line and performs classification in this one-dimensional space. The projection maximizes the distance between the means of the two classes while minimizing the variance within each class. This defines the Fisher criterion, which is maximized over all linear projections, w:

$$J(w) = \frac{|m_1 - m_2|^2}{s_1^2 + s_2^2} \qquad \text{Eq. 2.2}$$

Where m represents a mean, s^2 represents a variance, and the subscripts denote the two classes. In signal theory, this criterion is also known as the signal-to-interference ratio. Maximizing this criterion yields a closed form solution that involves the inverse of a covariance-like matrix. This method has strong parallels to linear perceptions. We learn the threshold by optimizing a cost function on the training set.

2.2 Support Vector Machine

It is required that each data instance is represented as a vector of real numbers for SVM. If there is a categorical attributes, we first have to convert them into numeric data. Also some experiences indicate that if the number of values in an attribute is too many, it might be more stable than using a single number to represent a

categorical attribute. It is also recommended to avoid attributes in greater numeric ranges. Scaling them before applying SVM is very important.

There are some kernels that can be used for SVM. We have to decide which one to use. Then the penalty parameter *C* and kernel parameters are chosen. As a kernel, RBF is a reasonable choice. The RBF kernel nonlinearly maps samples into a higher dimensional space, so unlike the linear kernel, it can handle the case attribute relations are nonlinear.

SVMs are motivated by many of the same considerations with linear machines, but rely on preprocessing the data to represent patterns in a high dimension. The goal in training a SVM is to find the separating hyper plane with the largest margin, and to find the weight vector **a** that maximizes *b*. (Duda, Hart & Stork 2001)

$$\frac{z_k g(y_k)}{\|a\|} \geq b \qquad k=1,\ldots,n; \qquad \text{Eq. 2.3}$$

If N_s denotes the total number of support vectors, then for n training patterns the expected value of the generalization error rate is bounded, according to

$$\varepsilon_n[error.rate] \leq \frac{\varepsilon_n[N_s]}{n}, \qquad \text{Eq. 2.4}$$

where the expectation is over all training sets of size n drawn from the distributions describing the categories.

2.3 *k*-Nearest-Neighbor Rule

With k-nearest neighborhood algorithm (KNN), classification is made by a majority of its neighbors. This algorithm is very simple. Object is assigned to the class which is the most common between its k nearest neighbors. It's better to choose an odd number as k to avoid equal numbers of classes. Also, choice of k depends on

data. For example, larger values of k reduce the noise effect on classification and may cause different results as seen in Figure 2.2. Some different techniques can be used to select most suitable k value. Cross-validation is one of them. In some special cases, when k=1, training sample is called nearest neighbor algorithm.

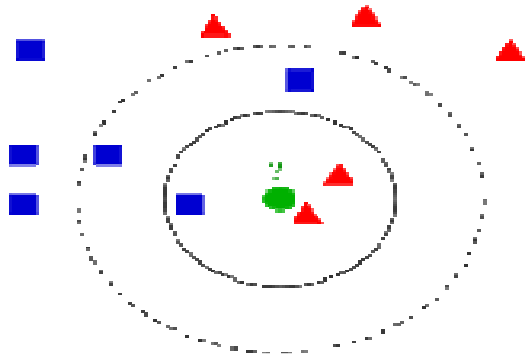

Figure 2.2 The test sample can be classified differently depends on the value of k. It can either be red triangle or blue squares.

2.4 Neural Networks

Neural Networks are electronic networks that process records. This term is used to refer a network or circuit of abstract neurons. Neural Networks are made up for connecting neurons and programming constructs that mimic like biological neurons. This can be explained like mathematical modeling of the neural systems. In more practical terms neural Networks are non-linear statistical data modeling or decision making tools. They can be used to made complex relationship between inputs and outputs or to find patterns in data. An artificial neural network includes simple processing elements which work as determined by the connections and element parameters. The best utility of the neural networks is that they can be used to infer a function from data, and also to use it. Basically, they implement linear discriminants where the inputs have been mapped nonlinearly. This is very useful if the data is complex.

We can talk about three parts of layers in neural networks as seen in figure 2.3. First one is input layer which consists of values in data that constitute inputs to the next layer of neurons. The next part of layers is called a hidden layer. This part may have several hidden layers of neurons. The final layer is the output layer where there is only one node for each class and the record is assigned to whichever class' node had the highest value.

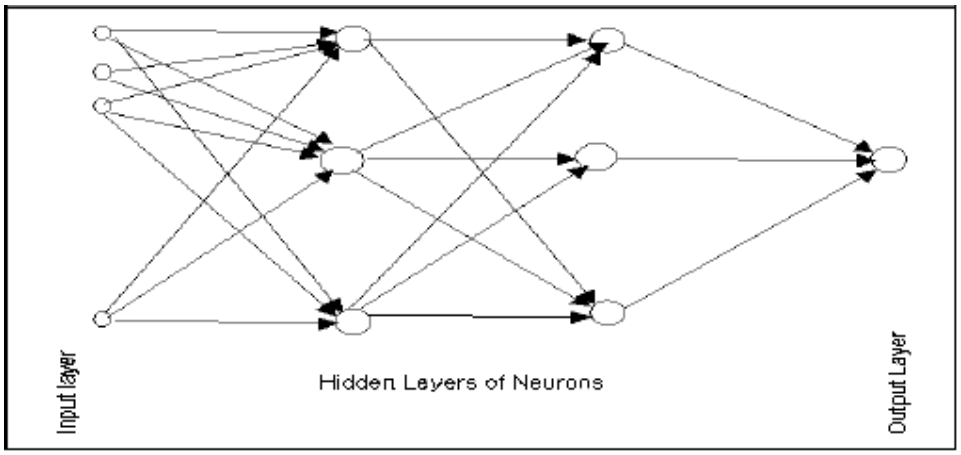

Figure 2.3 Three Layer neural network

Each input vector is presented to the input layer, and the output equals the corresponding component in the vector. Hidden units create their net *(net activation)* from these inputs. Net activation is the inner product of the inputs with the weights at the hidden unit:

$$net_j = \sum_{i=1}^{d} x_i w_{ji} + w_{j0} = \sum_{i=0}^{d} x_i w_{ji} \equiv W_j^t X, \qquad \text{Eq. 2.5}$$

In this Formula, i is the number of units on the input layer, j is for hidden layer. w_{ji} denotes the input-to-hidden layer weights at the hidden unit j. These hidden units create outputs which are simply nonlinear function of their activation,

$$y_j = f(net_j) \qquad \text{Eq. 2.6}$$

This is a simple threshold or sign function as an example for y_j

$$f(net) = Sgn(net) \equiv \begin{cases} 1 & if, net \geq 0 \\ -1 & if, net < 0 \end{cases} \qquad \text{Eq. 2.7}$$

CHAPTER THREE
DATASET & CLASSIFICATION RESULTS

In this chapter, we examine our database in more detail. Explanations of all numerical values, which mean different attributes of breast cancer samples are given.

Our classification methods are compared with confusion matrices. Best classification method is determined for our dataset.

3.1 Dataset

In this study, the database is breast cancer database which was obtained from University of Wiscounsin Hospitals Madison (Wolberg 1991). There are 11 attributes for each sample (Table 3.1). Attributes 2 through 10 have been used to represent instances respectively. Number of instances is 699. But some of the instances are deleted due to missing attributes. There is a class attribute in addition to 9 other attributes. Each instance has one of 2 possibilities: Benign or malignant. One of the other numeric value columns is ID column of instances. We pick this column out.

Our dataset includes two classes as we said. They are *benign (B)* and *malignant (M).* Numbers of the instances that belongs to these classes are not equal. So we may use a limited number of instances that are in benign class.

Table 3.1 Breast cancer database

#	Attribute	Domain
1	Sample code number	id number
2	Clump Thickness	1 – 10
3	Uniformity of Cell Size	1 – 10
4	Uniformity of Cell Shape	1 – 10
5	Marginal Adhesion	1 – 10
6	Single Epithelial Cell Size	1 – 10
7	Bare Nuclei	1 – 10
8	Bland Chromatin	1 – 10
9	Normal Nucleoli	1 – 10
10	Mitoses	1 – 10
11	Class	2(B), 4(M)

3.2 Classification Results

There are 4 classification methods in this study. Among them are linear discriminant, support vector machine, neural network and k-nearest neighborhood rule. All methods have different accuracy depending on its appropriateness to dataset.

For having a classification, database has to be separated to two parts. They are train and test datasets. Train dataset is for training our classification method. Later, we will try classification on test database by this trained classification method.

Benign instances are much more than malignant instances. So, for having equal instances, we use only 240 of our benign instances. This is equal to total number of malignant instances. 230 of these instances are for train data, 10 of these instances are for test data. Shortly, we have 460 instances for training, 20 instances for testing. Both of these datasets have equal number of instances from each class, B or M. Figure 3.1 shows this allocation.

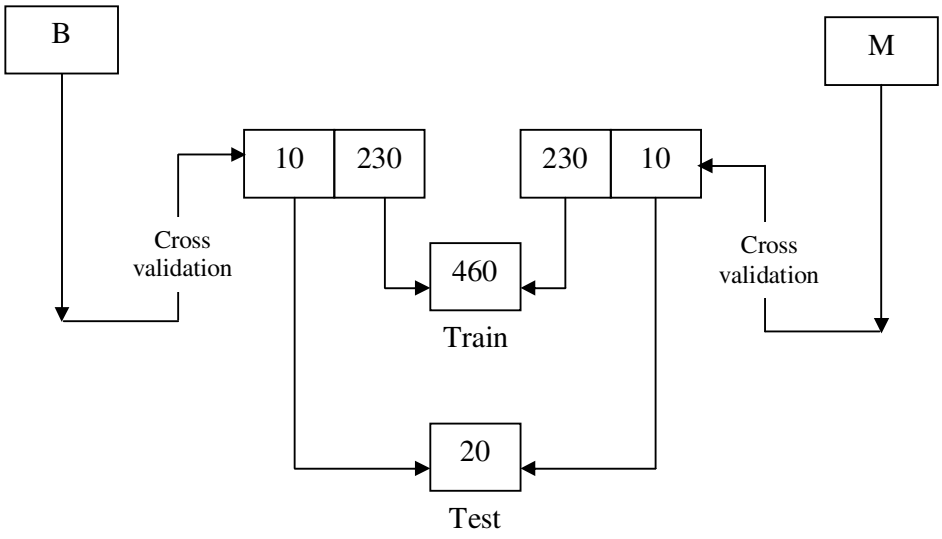

Figure 3.1 Choosing train and test dataset

Cross validation (Duda, Hart & Stork 2001), is used for all methods. The purpose of using cross validation is to have an average result for every calculation. By this way, we are having more realistic value and we easily avoid different peak results of our classification. Cross validation is repeated for choosing test dataset before every classification process.

There are four possible outcomes when classifying 2 data groups. These are true decisions of malignant and benign data and wrong decision of malignant and benign data. When we create confusion matrix as seen in Table 3.2 (Duda, Hart & Stork 2001), this matrix has these parts, giving us idea about success of classification.

Let's call these four possibilities A, B, C and D. A means malignant sample classified as malignant. B means malignant samples classified as benign. C means benign samples classified as benign. D means benign samples classified as malignant. Shortly, A and C are our success percentage, B and D are our error percentage. Overall accuracy is calculated simply taking average of A and C if train data numbers are the same for both classes.

Table 3.2 Confusion matrix

	Benign (classification result)	Malignant (classification result)
Benign (real result)	C	D
Malignant (real result)	B	A

Our confusion matrices are directly giving the percentage of classifications. In some studies, these A, B, C and D values are sample numbers. Then two terms are most commonly used. These are

$$Sensitivity = \frac{A}{A+B} \qquad \text{Eq.3.1}$$

$$Specifity = \frac{C}{C+D} \qquad \text{Eq.3.2}$$

These sensitivity and specifity values are given in confusion matrices in the following chapters.

3.2.1 Linear Discriminant Results

Table 3.3 shows accuracy percentage of linear discrimination method. As we have 10 benign and 10 malignant test data, minimum and maximum values are 85% and 100%. Over all accuracy isn't that integer because this value is calculated after 50 cross-validations. Average value is simply calculated by dividing total of accuracy values to validation number.

Linear discriminant functions are more successful at benign samples than malignant samples. Average accuracy is linear average of these two results, Malignant-Malignant (M-M) and Benign-Benign (B-B) as sample numbers are equal.

It is possible to say that linear classification method is the most suitable one between our methods for benign instance, because it has the best accuracy, 98%.

Table 3.3 Minimum, maximum and overall accuracy of linear discrimination

Cross-validated Original		Predicted Group Membership		Total %
		Benign %	Malignant %	
Min. accuracy 85%	Benign	90	10	100
	Malignant	20	80	100
Max accuracy 100%	Benign	100	0	100
	Malignant	0	100	100
Overall accuracy 96.60%	Benign	98	2	100
	Malignant	4.8	95.2	100

3.2.2 SVM

Before SVM classification, the cross-validation is done again to prevent different peak values.

Apart from this, SVM method has some variables that affect the results. These variables are argument and C variables. If we don't determine these variables, Matlab® determines them automatically but different values of these variables may cause different results in the same database. The best way to make these variable values more clear is to calculate accuracies for the same database, with different *"arg"* and *"C"*. Table 3.4 shows our classification results for the same train and test data. First horizontal line shows *"C"* variable. First Column shows *"arg"* variable. As we can see from the table, most of the accuracies are the same. But when we suppose that C=6 and arg=13, and make some cross-validation to get average value; we get the best result depending on our accuracy table.

So we began our process with these constant values, later carried on with cross validation, and accuracy calculation.

Table 3.4 Choosing argument and C constant for SVM classification

	1	2	3	4	5	6	7
1	0.0600	0.0600	0.0600	0.0600	0.0600	0.0600	0.0600
2	0.0600	0.0600	0.0600	0.0600	0.0600	0.0600	0.0600
3	0.0600	0.0600	0.0600	0.0600	0.0600	0.0600	0.0600
4	0.0600	0.0600	0.0600	0.0600	0.0600	0.0600	0.0600
5	0.0600	0.0600	0.0600	0.0600	0.0600	0.0600	0.0600
6	0.0600	0.0600	0.0600	0.0600	0.0600	0.0600	0.0600
7	0.0600	0.0600	0.0600	0.0600	0.0600	0.0600	0.0600
8	0.0600	0.0600	0.0600	0.0600	0.0600	0.0600	0.0600
9	0.0600	0.0600	0.0600	0.0600	0.0600	0.0600	0.0600
10	0.0600	0.0600	0.0600	0.0600	0.0600	0.0600	0.0600
11	0.0600	0.0500	0.0500	0.0500	0.0500	0.0500	0.0500
12	0.0400	0.0300	0.0300	0.0300	0.0300	0.0300	0.0300
13	0.0200	0.0600	0.0500	0.0300	0.0300	*0.0200*	0.0400
14	0.0600	0.0600	0.0600	0.0600	0.0600	0.0500	0.0600
15	0.0600	0.0600	0.0600	0.0600	0.0500	0.0600	0.0600

After 50 calculations and having these average values, we see that overall accuracy of SVM method is 97.30. We have 97.20% benign to benign accuracy and 97.4% malignant to malignant accuracy. Malignant instances are classified more successfully than benign instances as seen in table 3.5.

Table 3.5 Minimum, maximum and overall accuracy of SVM

Cross-validated Original C=6, arg.=13, cross val.=50		Predicted Group Membership		Total %
		Benign %	Malignant %	
Min. accuracy 85%	Benign	80	20	100
	Malignant	90	10	100
Max accuracy 100%	Benign	100	0	100
	Malignant	0	100	100
Overall accuracy 97.30%	Benign	97.20	2.80	100
	Malignant	2.6	97.4	100

3.2.3 k-Nearest Neighborhood Rule

Table 3.6 Minimum, maximum and overall accuracy of k-NN

Cross-validated Original		Predicted Group Membership		Total %
		Benign %	Malignant %	
Min. accuracy 85%	Benign	80	20	100
	Malignant	90	10	100
Max accuracy 100%	Benign	100	0	100
	Malignant	0	100	100
Overall accuracy 97.20%	Benign	96.20	3.80	100
	Malignant	1.80	98.20	100

We again have the results of 50 calculations and have these average values; our overall accuracy of k-NN method is 97.20%. This result is very close to SVM. We have 96.20% benign to benign accuracy and 98.2% malignant to malignant accuracy. Malignant instances are classified much more successfully than benign instances. Minimum and maximum accuracies are the same. It is possible to say that k-NN classification method is the most suitable one between our methods for malignant instance, because it has the best accuracy, 98.2%.

3.2.4 Neural Network

Neural network classification is another method which is more successful at malignant instances as we see in table 3.7. Overall accuracy is 96.06%. Because of having different numbers of train data, overall accuracy is not average of malignant samples and benign samples.

Neural network classification is available for modification, and for having different algorithms like putting attributes in an order depending on importance.

Table 3.7 Minimum, maximum and overall accuracy of NN

Cross-validated Original		Predicted Group Membership		Total %
		Benign %	Malignant %	
Min. accuracy 93.99%	Benign	94.67	5.33	100
	Malignant	10.26	89.74	100
Max accuracy 97.87%	Benign	97.95	2.05	100
	Malignant	0	100	100
Overall accuracy 96.06%	Benign	95.96	4.04	100
	Malignant	3.34	96.66	100

3.3 Comparing Results

Figure 3.2 shows classification results. Depending on these numbers, it can be said that SVM is the most suitable classification method for our breast cancer dataset. SVM has biggest overall value: 97.3% Linear Discriminant classification has the biggest value with benign instances, as well as k-NN has with malignant instances. For breast cancer data, if we consider that we don't know the class label of a new instance, we easily can say that support vector machine classification will be the most feasible method to use.

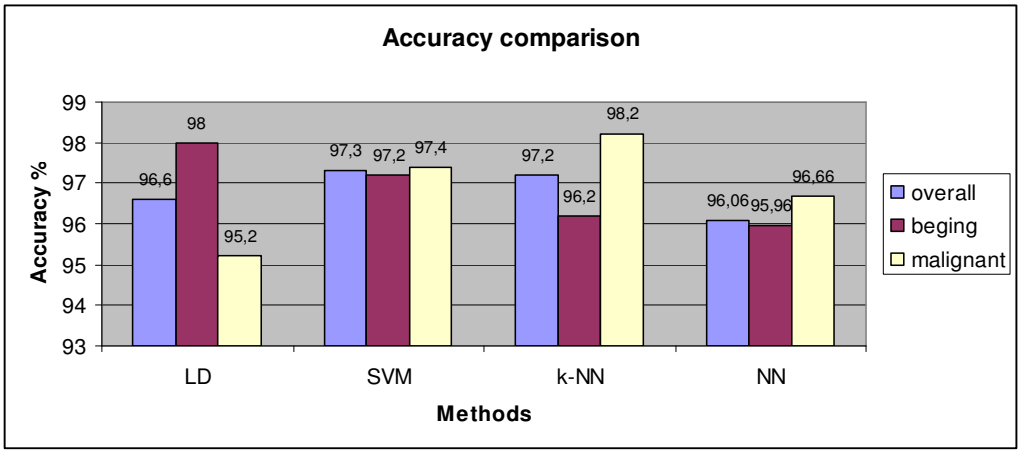

Figure 3.2 Accuracy comparison of methods

3.4 Comparing Results with Other Studies

Table 3.9 shows some classification results for both breast cancer database and other datasets. Many studies are based on having a classification result and developing new algorithms for increasing accuracy. As we see from the table, results and considerations are not different than other studies.

Support vector machine classification is the most widely used classification method for breast cancer dataset. It is suitable for structure of attributes, so it gives the best result. Linear discriminant classification is another easy to apply method for this dataset. It gives a valid result but its not as good as SVM. Neural network methods are as good as SVM but these methods need modification.

Mihir Sewak used SVM for Wisconsin Breast Cancer Database and got 99.20% accuracy result. I. Anagnostopoilosi had 97.90% accuracy by using advanced neural network techniques, as well as Z. Yang had 98.50%. These changes with the same methods depend on network specifications, importance order of attributes and some other reasons. Y.Li, T. Mu, W. Wolberg made a classification application for Wisconsin Breast Cancer Database. They all used support vector machine algorithm and they got 95.50%, 98.40% and 97.50% accuracies respectively.

If we have a look at our results, we can notice that our best accuracy is from support vector machine classification which is used very often for breast cancer data. Overall accuracy, 97.30% is near the average value of other studies. If we didn't use the same number of test data for both classes, we could get different accuracy values. Because as we mentioned in Ch.3.3; support vector machine method is more successful with malignant instances. Our neural network classification has an average accuracy percentage between other results. K nearest neighborhood rule is not a method that is used often for this dataset. Our results give us a hint about it; because it has the least accuracy percentage between other results. And it hasn't been used in other studies. 96.06% is the result of kNN.

Table 3.8 Classification results of some other studies with Wisconsin breast cancer database

	Test data	Class	Linear Disc.	SVM	NN	KNN
Mihir Sewak	57	2		99.29%		
Anagnostopoilosi, I.	569	2			97.90%	
Yang, Z.	263	2			98.50%	
Li, Y.	569	2		95.60%		
Mu, T.	10	2		98.40%		
Wolberg, W.,H.,	569	2		97.50%		
Güneşer, C	20	2	96.60%	97.30%	97.20%	96.06%

CHAPTER FOUR
CONCLUSIONS AND FUTURE WORK

Breast cancer is the second most leading cause of cancer death in women. In 2008 it caused 519000 deaths worldwide. Timing is very important to identify the type of cancer for reducing risk of death. Nothing can replace manual human diagnosis but automating the prediction process before reaching a result can be achieved with acceptable accuracy.

This study aimed to create a view for diagnosis with computer supported software programs and different classification algorithms. First of all, we have to obtain attributes' values in numerical values. Process until getting these values is totally medical. After having these attributes numerically, it is easy to write them down for using in different ways. We have to consider that each of these attribute values have different importance to have a result. They all affect the diagnosis.

Four classification methods are explained in Ch.2. Despite of the fact that the most suitable classification method is support vector machine classifier for this dataset, having results with all these four methods helps us to have an opinion. These classification methods are given from the most basic one to the most complex one respectively. Every dataset and every sickness attribute values have different accuracy with all these methods. It will be useful to decide which classification method to develop after getting average values with different methods. In every method, accuracy will change also depending on the class numbers.

In software screen, it's possible to see every single detail of classification process. There is chance to see which instance is classified correct, which instance is classified incorrect. We can take these incorrectly classified instances and we can examine them for finding reasons of incorrect classification. This means another block of software codes. These results can be used as feedbacks to classification algorithm. By these feedbacks, parameters can be changed automatically. Average value can be calculated and if accuracy results are better, these new values of

parameters are saved. This means the improvement of classification methods. With every mistake and every feedback, more ideal parameters can be achieved.

Our classification algorithms can be embedded in hospital automation programs. For instances that are hard to determine, these classification results can support manual diagnosis. With a simple graphical user interface, these methods can be used to have an idea about instances. Future studies include improving algorithms, creating an artificial proofreader and making these classification methods possible to use with user friendly diagnosis programs.

REFERENCES

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern Classification (2nd ed.).* NY: John Wiley & Sons.

Hsu, C. W., Chang, C. C., & Lin, C.J. (2008). *A practical guide to support vector classification.* Department of Computer Science, National Taiwan University.

Li, Y., Hu, Z., Cai, Y., & Zhang, W. (2005). *Support vector based prototype selection method for nearest neighbor rules.* Lecture Notes in Computer Science (3610), 528-535.

Mu, T., & Nandi, A. (2007). Breast cancer detection from FNA using SVM with different parameter tuning systems and SOM-RBF classifier. *Journal of the Franklin Institute,* (344), 285-311.

Sewak, M., Vaidya, P., Chan, C. C., & Duan, Z. H. (2007). *SVM approach to breast cancer classification.* 2nd International Multi Symposium on Computer and Computational Sciences 2007.

World Health Organization. (February, 2009). *Cancer Fact Sheet,* Retrieved April, 2009, from http://www.who.int/mediacentre/factsheets/fs297/en.

Wolberg, W. H., Street, W. N., & Mangasarian O. L. (1993). Image analysis and machine learning applied to breast cancer diagnosis and prognosis. *Analyt. Quant. Cytol and Histol.* (15), 396-404.

Wolberg, W. H. (1991). *Wisconsin breast cancer database,* University of Wisconsin Hospitals, Madison .

Xiong, X., Kim, Y., Beak, Y., Rhee, D. W., & Kim S.H. (2005*). Analysis of breast cancer using data mining & statistical techniques.* 6th Int. Conference on Software Engineering, AI (SNPD/SAWN'05).

Yang, H., & Pizzi, N. J. (2004). *Biomedical data classification using hierarchical clustering.* Dept. of Computer Science, University of Manitoba, Canada.

Yang, Z. R., Lu, W., Yu, D., & Harrison, R. G. (2000). *Detecting false benign breast cancer diagnosis.* Neural Networks, Proceedings of the IEEE-INNS-ENNS International Joint Conference, (3), 665-658.