Department of Computer Science and Engineering
University of Texas at Arlington
Arlington, TX 76019

Evidence Combination in Medical Data Mining

Gauri Mahajani, Y. Alp Aslandogan

{mahajani,alp}@cse.uta.edu

Technical Report CSE-2003-23

Evidence Combination in Medical Data Mining

Gauri Mahajani, Y. Alp Aslandogan
Department of Computer Science and Engineering
The University of Texas at Arlington, U.S.A.
{mahajani,alp}@cse.uta.edu

ABSTRACT: In this work we apply Dempster-Shafer's theory of evidence combination for mining medical data. We consider the classification task in two domains: breast tumors and skin lesions. Classifier outputs are used as a basis for computing beliefs. Dynamic uncertainty assessment is based on class differentiation. We combine the beliefs of three classifiers: k-Nearest Neighbor (kNN), Naïve Bayesian and Decision Tree. Dempster's rule of combination combines three beliefs to arrive at one final decision. Our experiments with k-fold cross validation shows that the nature of the data set has a bigger impact on some classifiers than others and the classification based on combined belief shows better accuracy than any individual classifier. We compare the performance of Dempster's combination with differentiation-based uncertainty assignment with those of linear and majority vote combination models. We study the circumstances under which the evidence combination approach improves classification.

Introduction

The aim of this work is to study and apply a formal evidence combination technique for mining medical data for prediction of or screening for a disease. Input data, consisting of feature vectors, is input to three different classifiers. The classifiers we used in this study are kNN (k nearest neighbor) [9], Bayesian [10] and Decision Tree classifier [10]. Each classifier provides beliefs for each class.

These evidences are then combined to reach a final diagnosis using Dempster's belief combination formula. The experiments are carried on the dermatology and breast cancer data. The comparative results and their explanations are provided.

Input Data Source

Input datasets used for evidence combination in medical data mining are the dermatology data and breast cancer data made available from University of California, Irvine machine learning repository at www.ics.uci.edu/~mlearn/MLRepository.html. Dataset of breast cancer consists of nine features and a classification of each record. The classification can be done in two different classes namely benign and malignant. Dermatology dataset consists of thirty-four features. Each record can be classified as one of the six classes. Classes are namely psoriasis, seboreic dermatitis, lichen planus, pityriasis rosea, cronic dermatitis, and pityriasis rubra pilaris.

Dataset of breast cancer cases consists of 682 records whereas dermatology dataset consists of 366 records. Training data set contains feature values as well as classification of each record. Testing is done by k-fold cross validation method. Dataset was divided into training and testing set by choosing one-fourth records as test cases.

These test records were not used for training purpose. Testing was carried out until every record appeared in the test set. A confusion matrix is computed for every test.

Introduction to Belief Functions, Evidence Theory

Belief is a trust or confidence. Let us consider that event A occurred and let there be an evidence available providing a degree of support for the occurrence of event A. Combining all degrees of support for event A forms a numerical measure of belief that event A occurred. Thus belief in this context formalizes confidence for the occurrence of event A. A mathematical function that translates degree of support to belief is known as Belief function. It assigns degrees of confidence to the events.

Properties of basic belief $m(X)$ are as follows

1. $\sum_{X \in \Omega} m(X) = 1$

2. $m(\phi) = 0$ where ϕ is empty. This indicates belief of empty set is always zero.

Belief function for an event A can be $\mathrm{Bel}(A) = \sum_{X \subseteq A \text{ and } A \in \Omega} m(X)$
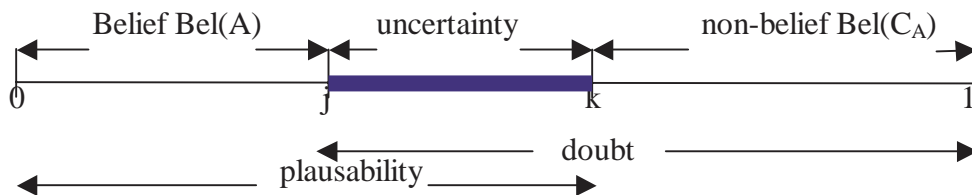

Figure 2.1 Different Measurements Over a Unit Interval.

Description of intervals is as follows:

- interval 0-j indicates belief, Bel(A), in event A

- interval k-1 indicates non-belief. This is a negation of belief, Bel(C_A). C_A stands for compliment of event A.

- interval j-k indicates uncertainty which means that neither belief nor non-belief intervals are chosen.

- Plausability, pl(A) is the sum of belief and uncertainty measure. This indicates that maximum extension of belief lies in the uncertainty, which neither proves belief nor disproves it.

- Doubt is the sum of non-belief and uncertainty measure.

Following are the properties for the above measures:

$$\text{Bel}(A) \leq \text{pl}(A) \qquad \text{Bel}(A) + \text{Bel}(C_A) \leq 1$$

$$\text{pl}(A) = 1 - \text{Bel}(C_A) \qquad \text{pl}(A) + \text{pl}(C_A) \geq 1$$

Theory of evidence deals with the evaluation of beliefs from a number of evidences and their combination. For example consider three sources of evidence named E1, E2 and E3. Let the event space be $\Omega = \{A, B, C\}$. Evidences provide measures for the event space. These measures include belief for each event and uncertainty. Thus measures assigned by evidence E1 are given as $\text{Bel}_{E1}(A)$, $\text{Bel}_{E1}(B)$, $\text{Bel}_{E1}(C)$ and $\text{Bel}_{E1}(\text{uncertainty})$. Note that $\text{Bel}_{E1}(A) + \text{Bel}_{E1}(B) + \text{Bel}_{E1}(C) + \text{Bel}_{E1}(\text{uncertainty}) = 1$. Similarly for E2 and E3 measures are $\text{Bel}_{E2}(A)$, $\text{Bel}_{E2}(B)$, $\text{Bel}_{E2}(C)$, $\text{Bel}_{E2}(\text{uncertainty})$ and $\text{Bel}_{E3}(A)$, $\text{Bel}_{E3}(B)$, $\text{Bel}_{E3}(C)$ and $\text{Bel}_{E3}(\text{uncertainty})$. A decision can be made based on a combination of these beliefs.

In this research we use classifier output to form evidence and a decision of benign or malignant forms an event. Thus Ω = {benign, malignant}. The sources of evidence are k-nearest neighbor classifier, a Naive Bayesian classifier and a Decision tree classifier. Evidence assigns basic belief masses to the events and also a degree of uncertainty. These measures are then combined to arrive at the final classification.

kNN Classification

kNN classifier stands for k nearest neighbor classifier, where k is a user specified parameter indicating a number of neighbors to be considered. The input to the classifier is a set of features. The output is a classification of the sample for example benign or malignant. After reading input features kNN sorts each feature vector. The classifier finds 'k' nearest neighbors for each test input feature. Identification of the nearest neighbor is based on the distance between test input feature value and the sorted training feature values. Training inputs corresponding to first k shortest distances are chosen as k nearest neighbors. Let us consider an example having a feature set of four. Let k = 2 and there are two classes. Since every feature gets two neighbors in all eight neighbors are evaluated. We already know the classification of neighbors. kNN considers classification of neighbors as a vote. The following table explains the kNN classification process.

Belief evaluation from kNN Classifier Results

As explained in the previous section, kNN considers nearest neighbors as voters. The distance measures evaluated from these neighbors are used to compute beliefs for classes. Distance between test case feature value and neighbor feature value is

calculated. Let us denote this distance by 'd$_s$'. This distance is normalized in the range 0 to1. A fraction of this distance is calculated as d$_s$/d$_{mean}$ where, d$_{mean}$ is the average distance among the samples belonging to same class and is normalized into the range 0 to 1. Fraction would be greater than 1 if the distance of the test case attribute is more than the average distance for the class and less than 1 otherwise.

To evaluate a distance measure a decreasing function of the distance d$_s$ must be applied. The reason behind this is that as distance between the test case feature and its neighbor feature decreases the possibility that two cases belong to same class increases and the confidence in the event of test sample belonging to same class as the neighbor also increases.

The following distance function is used,

$$\text{Distance measure} = e^{-\frac{ds}{dmean}} \quad [19]$$

$$\text{where} \quad e^{-\frac{ds}{dmean}} = 1 \text{ when ds} = 0$$

and

$$\lim_{ds \to \infty} e^{-\frac{ds}{dmean}} = 0$$

Thus belief mass of a class is the average of all such distance measures voting for that class. Belief masses for the classes are then normalized so that

$$\sum_{i=1}^{K} m(i) = 1$$

Bayesian Classification

This classifier is based on Bayesian probability principle. The classifier takes feature vector as an input. Given the training set, it computes the apriori conditional probabilities. An apriori conditional probability indicates the probability of occurrence of a feature value given a class label, represented as P(Feature value | Class). According to Bayesian probability theory posterior probability of the class is a conditional probability of class given feature values. It is denoted as P(Class | Feature set). Bayesian probability theory states that knowing conditional probability of feature set values and prior sample probability of a class, posterior probability can be evaluated. Thus we get,

$$P(class \mid feature\,set) = \frac{P(feature\,set \mid class) \times P(class)}{P(feature\,set)}$$

where *P(feature set / class)* is a conditional probability of occurrence of feature set for a given class, *P(class)* is a prior sample probability of occurrence of the class and P(feature set) is a probability of occurrence of a feature set. Let us consider that feature set contains two features. Thus *P(feature set / class)* will be *P(f1∩f2 / class)*. In case of dermatology and breast cancer data, features are independent of each other. Thus *P(f1∩f2 / class)* becomes *P(f1/class) X P(f2/class)* . Similarly *P(f1 ∩ f2)* will be P(f1) x P(f2). The classifier computes conditional probabilities for each feature beforehand. Thus posterior probability is calculated as follows.

$$P(class \mid feature\ set) = \frac{P(f1 \mid class) \times P(f2 \mid class) \times P(class)}{P(f1) \times P(f2)}$$

Using these values Bayesian classifier calculates posterior probability for each class. These probabilities form a basis of decision. The class with largest posterior probability will be the final decision. These normalized posterior probabilities are used to evaluate basic beliefs of the Bayesian classifier.

Decision Tree Classification

This classifier uses decision tree evaluation for classification. Input to the classifier is same as other classifiers with feature values and classification for each record. From this dataset, evaluates a decision tree. Leafs are given class labels, branches of the tree are conjuncts forming the rule. Conjuncts are the 'splits' at attribute value. At each node a test attribute is chosen according to Information gain measure. In the classification mode classifier starts at the root node and follows the branches. When it reaches leaf, the class specified at the leaf is the final classification. Association rules are then extracted from the decision tree. An association rule is of type A → B. It means A suggests or implies B. A, B can form a single item or a set of items

Belief Evaluation from Decision Tree Classification Results

As explained in the previous section the decision tree classifier builds a decision tree. Association rules can be extracted from this tree. An association rule has support and confidence associated with it.

$$Support = \frac{Number\ of\ records\ with\ A\ and\ B}{Total\ number\ of\ records}$$

where numerator indicates the number of records with A and B both true.

$$Confidence = \frac{Number \; of \; records \; with \; A \; and \; B}{Total \; number \; of \; records \; with \; A}$$

Confidence can also be written as $\frac{P(A \cap B)}{P(A)}$

where,

P(A ∩ B) is the probability of A ∩ B.

In case of our dataset classification process $P(A \cap B)$ forms the probability of the occurrence of feature values with given class. $P(A)$ indicates probability of A that is the probability of features .

But,

P(A ∩ B) = P(A|B) x P(B).

Thus,

$$Confidence = \frac{P(feature \; set \mid class) \times P(class)}{P(feature \; set)}$$

which is the posterior probability of a class for given feature set. Thus confidence can be used to form basic belief of the decision tree classifier.

Uncertainty Evaluation

So far we have discussed each classifier and its basis of classification. This section explains how to evaluate uncertainty for each classifier. Distance values in kNN can be used for evaluation of belief masses. Similarly posterior probability in Bayesian classifier and confidence measure in decision tree classifier can be used to evaluate belief masses.

Next step is evaluation of uncertainty according to the belief masses. The idea behind uncertainty is that the closer the values of beliefs for K classes to each other, the more uncertain the classifier is about its decision. As the beliefs start spreading apart uncertainty starts decreasing. Let uncertainty be denoted as H(U) [3]

Idea behind the uncertainty evaluation is if the number of classes is K then the distance between the belief value and the value 1/K is evaluated. If all the classes have the same distance then the ambiguity involved in the classification is higher. If one class shows maximum distance then the ambiguity involved is less.

Uncertainty is given as

$$H(U) = 1 - \frac{K}{K-1} \sum_{i=1}^{K} (m(i) - \frac{1}{K})^2 \qquad [3]$$

Thus the belief is recalculated as Bel(i) = α m(i) and uncertainty as Bel(θ)= β H(U).

$$\sum_{i=1}^{K} Bel(i) + Bel(\theta) = \sum_{i=1}^{K} \alpha \, m(i) + \beta \, H(U)$$

But $$\sum_{i=1}^{K} Bel(i) + Bel(\theta) = 1$$ and $$\sum_{i=1}^{K} m(i) = 1$$

Thus value of α = 1- β H(U). Value of β is chosen as 0.3 after checking performance results for various values in the range of 0 to 1.

According to our observations, Bayesian classifier results are biased towards the class having larger prior sample probability. In such cases uncertainty is adjusted

accordingly. When the class with highest belief has more prior sample probability than the class with the next highest belief then the uncertainty is calculated using different expression. Let the class with maximum belief has prior sample probability of $p1$ and the class with the second highest belief has the prior sample probability be $p2$. The uncertainty is evaluated by following expression $0.35 \times e^{p1/p2}$. The value of 0.35 is the ratio of failures when highest belief has more prior class probability than the next highest belief to the total number of failures. This ratio indicates the percentage of uncertainty introduced in Bayesian classification due to such typical cases. The beliefs are then adjusted proportionally so that

$$\sum_{i=1}^{K} Bel(i) + \text{uncertainty} = 1$$

Overview of Dempster Shafer Approach

Consider a situation in which there are N different evidences providing decision for the same event. Evidences are arriving at decisions independently. Some evidences are more accurate while arriving at a decision in certain circumstances than other. In such cases when accurate decision of any category is important, relying on a single classifier may result in more misclassifications, which would be more costly. Dempster's rule of combination deals with such cases by combining the decisions to arrive at final decision.

Dempster-Shafer theory takes into consideration beliefs of evidences in the form of mathematically evaluated beliefs. Belief of evidence takes values in the range [0, 1]. Let events be a, b, c. Frame of discernment Ω is defined as {a, b, c}. Summation of all such beliefs over frame of discernment Ω is equal to 1. The entire set of events, given as θ, also gets a belief value. Thus belief function explicitly represents ambiguity or ignorance [4]. Thus $Bel(x) + Bel(\sim x) + Bel(\theta) = 1.0$.

Dempster's rule of combination deals with these beliefs. Rule assumes that observations are independent and have a non-empty set intersection [4]. Any two beliefs Bel1 and Bel2 with elements Ai and Bi respectively may be combined into a new belief function using Dempster's rule of combination [4]. Let combined belief mass is assigned to Ck, where C is a set of all subsets produced by A∩B. The mathematical representation of the rule is as follows:

$$Bel(Ck) = \frac{\sum_{Ai \cap Bi=Ck;Ck \neq \phi} Bel(Ai) \times Bel(Bi)}{1 - \sum_{Aj \cap Bj=\phi} Bel(Aj) \times Bel(Bj)}$$

Evidences considered here are the classifiers namely k-nearest neighbor, Bayesian and decision tree. The decision provided by these classifiers is the category of the cancer. Let's consider breast cancer data having nine features and two classes, benign and malignant. Each classifier provides belief for classes and uncertainty. These beliefs are then combined using Dempster's rule of combination.

Steps of Combination

First step of Dempster's combination involves evaluation of beliefs and uncertainties from individual classifiers along with classifier decision. Thus initial step evaluates decisions from evidences, which are later combined. The following figure shows the inputs and outputs of the individual classifiers.

Features   Features   Features

| Bayesian Classifier | k-Nearest Neighbor Classifier | Decision tree Classifier |

Beliefs for each class including uncertainty & Classification Beliefs for each class including uncertainty & Classification Beliefs for each class including uncertainty & Classification
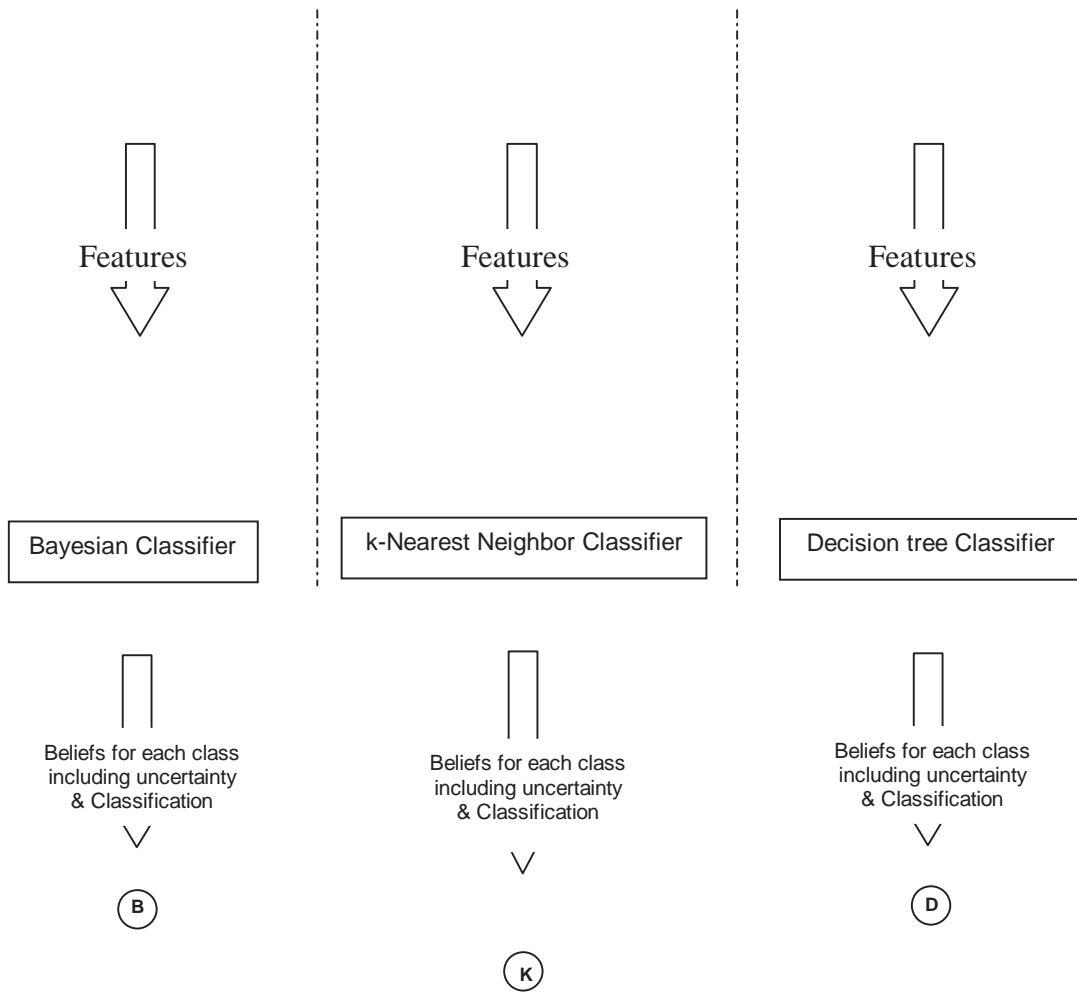
(B)   (K)   (D)

Figure 3.1 Block Diagram of Individual Classifiers

Output 'B' indicates belief values obtained from Bayesian classifier similarly 'K' and 'D' outputs are belief values from k-nearest neighbor and dcision tree classifier respectively.

The first step combines beliefs of k-nearest neighbor classifier and Bayesian classifier results. Dempster's rule of combination starts with the belief and uncertainty as inputs. It combines the pieces of evidences, which in combination form a same hypothesis. B and K are the evidences providing beliefs for benignn and malignant classes obtained from Bayesian and k-nearest neighbor classifiers respectively.
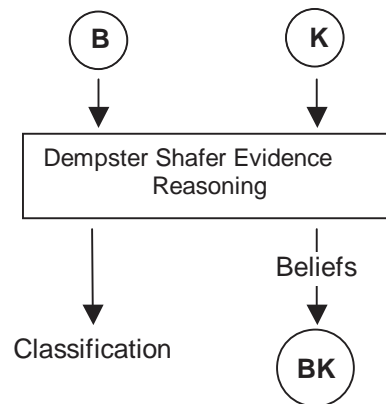

Figure 3.2 Combination of Bayes and kNN Classifier

Let kNN classifier provides beliefs Bel_kNN(B) and Bel_kNN(M), where Bel_kNN indicates belief provided by k-nearest neighbor and B, M are the two classes (benign and malignant) under consideration. Similarly for Bayesian classifier beliefs are given as Bel_Bayes(B) and Bel_Bayes(M). Uncertainties for two classifiers are U_kNN and U_Bayes respectively. Thus matrix under consideration is as follows

		Bayes Benign	Malignant	Uncertainty
kNN		Bel_Bayes(B)	Bel_Bayes(M)	U_Bayes
Benign	Bel_kNN(B)	Bel(B)	Bel(Ø)	Bel_kNN(B) x U_Bayes
Malignant	Bel_kNN(M)	Bel(Ø)	Bel(M)	Bel_kNN(M) x U_Bayes
Uncertainty	U_kNN	Bel_Bayes(B) x U_kNN	Bel _Bayes(M) x U_kNN	Bel(U)

Bel(B) is a belief mass given to class benign. As shown in the matrix it is evaluated by multiplying benign belief masses of kNN and Bayes, uncertainty for kNN and benign belief of kNN, benign belief of Bayes and uncertainty of kNN. While finally combining benign belief, all these basic beliefs are added. Thus,

$$Bel\_combin\,ed(B) = Bel\_Bayes(B) \times Bel\_kNN(B) + U\_Bayes \times Bel\_kNN(B) + Bel\_Bayes(B) \times U\_kNN$$

This combined belief is then normalized by factor $1 - \Sigma \, A \cap B$ (where $A \cap B = \Phi$).

Thus final combined belief is computed as,

$$Bel\_Comb(B) = \frac{Bel\_Bayes(B) \times Bel\_kNN(B) + U\_Bayes \times Bel\_kNN(B) + Bel\_Bayes(B) \times U\_kNN}{1 - Bel\_Bayes(M) \text{ x } Bel\_kNN(B) - Bel\_Bayes(B) \text{ x } Bel\_kNN(M)}$$

Similarly combined beliefs for malignant class and uncertainty are given as follows:

$$Bel\_Comb(M) = \frac{Bel\_Bayes(M) \text{ x } Bel\_kNN(M) + U\_Bayes \text{ x } Bel\_kNN(M) + Bel\_Bayes(M) \text{x } U\_kNN}{1 - Bel\_Bayes(M) \text{ x } Bel\_kNN(B) - Bel\_Bayes(B) \text{ x } Bel\_kNN(M)}$$

$$Bel\_Comb(U) = \frac{Bel(U)}{1 - Bel\_Bayes(M) \text{ x } Bel\_kNN(B) - Bel\_Bayes(B) \text{x} Bel\_kNN(M)}$$

The final diagnosis of this combined classifier is the class with highest combined belief. Similarly combination classifiers of kNN+decision tree and Bayes+decision tree are also studied. Following is the block diagram of combination of k-nearest neighbor, Bayes and decision tree classifier.
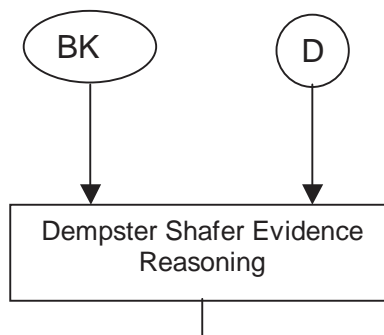
Classification

Figure 3.3 Combination of Bayes, kNN and Decision Tree Classifier

Mathematical Calculations:

$$Bel\_Comb(M) = \frac{Bel\_BK(M) \times Bel\_D(M) + U\_BK \times Bel\_D(M) + Bel\_BK(M) \times U\_D}{1 - Bel\_BK(M) \times Bel\_D(B) - Bel\_BK(B) \times Bel\_D(M)}$$

$$Bel\_Comb(B) = \frac{Bel\_BK(B) \times Bel\_D(B) + U\_BK \times Bel\_D(B) + Bel\_BK(B) \times U\_D}{1 - Bel\_BK(M) \times Bel\_D(B) - Bel\_BK(B) \times Bel\_D(M)}$$

$$Bel\_Comb(U) = \frac{Bel(U)}{1 - Bel\_BK(M) \times Bel\_D(B) - Bel\_BK(B) \times Bel\_D(M)}$$

Thus combination takes advantage of the fact that individual classifier is more accurate in classifying records corresponding to a certain class than the other. Thus after combination the overall classification becomes more accurate.

Experimental Setup and Methodology

Test results were carried out on dermatology dataset and breast cancer dataset. The explanation of the datasets used is as follows. The breast cancer data has total number of instances in this dataset is 682. Number of attributes is 10. Records belong to one of two classes. All the attributes take values between zero and ten. The classes are benign and malignant and they are denoted as 0 and 1 respectively. 444 records belong to class benign and 238 records belong to class malignant. The differential diagnosis of erythmato-sqamous is a real problem in dermatology. Types share the clinical features of erythema and scaling, with very little differences [13]. Dataset contains 34 attributes out of which 33 are linear and one is nominal. The diseases in this group are psoriasis, seboreic dermatitis, lichen planus, pityriasis rosea, cronic dermatitis, and pityriasis rubra pilaris. These diseases form the classes for classification. Dataset consists of 358 records. Classification into these classes is denoted as numbers from 0 to 5 respectively. Total records belonging to each class are 112, 61, 72, 49, 52, and 20 respectively. Distribution of datasets among classes is proportional and thus all the

classes are subdivided into 4 different subsets. For testing purpose one subset of each class is kept aside as a test case and the remaining three subsets are used for training purpose.

Test Data Results

This section provides the results of individual classifiers and a combination classifier. The results are provided in the form of a confusion matrix. Two sample result sets are provided for dermatology data and four result sets are provided for breast cancer data. After each test case result a small description about the result is provided.

Following are the test results of breast cancer dataset. Confusion matrices are listed for four classifiers namely k-nearest neighbor denoted as kNN, Bayesian denoted as Bayes, decision tree denoted as D-tree and combination classifier of kNN, Bayesian and decision tree denoted as kNN+Bayes+D-tree. The class denoted by 2 corresponds too uncertainty classification.

Table 4.1 Test Case 1

kNN				Bayes				D-Tree				kNN+Bayes+D-Tree			
	0	1	2		0	1	2		0	1	2		0	1	2
0:	105	2	0	0:	102	5	0	0:	104	3	0	0:	104	3	0
1:	15	48	0	1:	4	59	0	1:	13	49	0	1:	4	59	0

kNN classifier shows maximum accuracy in classification of records belonging to class 0. Bayesian classifier shows maximum accuracy in classification of records

belonging to class 1. Decision tree classifier is less accurate in classifying records of both the classes. As evident from the result set, combination classifier is the most accurate classifier. Similarly other tests are carried out

According to the results taken the accuracy calculation of the test cases is as follows

Table 4.5 Accuracy of Classifiers for Test Cases on Breast Cancer Data.

	kNN (%)	Bayesian (%)	Decision tree (%)	kNN + Bayes + D-Tree (%)
Test 1	90.0	94.7	90.0	95.8
Test 2	90.9	91.5	94.5	93.9
Test 3	90.7	96.5	90.7	95.9
Test 4	96.6	96.6	89.7	97.1

Overall accuracy of kNN classifier is 92%, Bayesian classifier is 93% and decision tree classifier is 91%. The combination classifier overall accuracy is 95.7%, which is improved over individual classifiers.

Following are the test results of dermatology dataset. Test case 1 consists of 91 test records. The class denoted by 6 corresponds too uncertainty classification.

Table 4.6 Dermatology Data Test Case 1 (kNN)

	0	1	2	3	4	5	6
0:	21	0	1	0	0	0	6
1:	0	0	14	0	0	0	1
2:	0	0	18	0	0	0	0
3:	0	0	12	0	0	0	0

4:	0	0	11	0	0	0	2
5:	0	0	0	0	0	0	5

Table 4.7 Dermatology Data Test Case 1 (Bayes)

	0	1	2	3	4	5	6
0:	28	0	0	0	0	0	0
1:	1	13	0	1	0	0	0
2:	0	0	18	0	0	0	0
3:	0	2	0	10	0	0	0
4:	0	0	0	0	13	0	0
5:	1	0	0	0	0	4	0

Table 4.8 Dermatology Data Test Case 1 (D-Tree)

	0	1	2	3	4	5	6
0:	28	0	0	0	0	0	0
1:	1	13	0	0	0	0	1
2:	0	0	18	0	0	0	0
3:	0	1	0	11	0	0	0
4:	0	0	0	0	13	0	0
5:	0	0	0	0	0	5	0

Table 4.9 Dermatology Data Test Case 1 (kNN+Bayes+D-Tree)

	0	1	2	3	4	5	6
0:	28	0	0	0	0	0	0
1:	0	15	0	0	0	0	0
2:	0	0	18	0	0	0	0
3:	0	0	0	12	0	0	0
4:	0	0	0	0	13	0	0
5:	0	0	0	0	0	5	0

The combination classifier improves the classification. It classifies all the records belonging to class 1 and class 3 and class 5 correctly which were misclassified by individual classifiers.

Similarly other tests are carried out and the accuracy results are as shown in the following table.

Table 4.14 Dermatology Data Test Accuracy Results

	kNN (%)	Bayesian (%)	Decision tree (%)	kNN + Bayes + D-Tree (%)
Test 1	43	94.5	96.7	100
Test 2	40.6	91.2	100	100
Test 3	47.2	95.6	94.5	94.5
Test 4	43.5	96.47	98.8	98.8
Test 5	52.8	97.7	89.65	98.8
Test 6	27.7	95.5	90.0	95.5

Overall accuracy of kNN classifier is 42.5%. Bayesian classifier shows overall accuracy of 95%. Decision tree classifier shows overall classification of 94.9%. Overall combination classifier accuracy is 97.9%. Thus combination shows improvement in the classification accuracy. This indicates that combination classifier is the best of all the classifiers. Different classifiers have varying performance on different datasets but the combination classifier shows overall steady performance.

<div align="center">Comparison with Majority Vote Fusion Technique</div>

One of the methods to combine classification of different classifiers is to take a majority vote. Each classifier is provided with the same training and testing files. During classification each classifier votes for a class. The fusion technique chooses the class with the highest vote. If none of the classifiers agrees on a single class then the final classification remains unknown or uncertain. This is the easiest way of classifier fusion and requires no complex calculations.

Above mentioned majority vote technique is compared with the Dempster Shafer combination results. Following are the results obtained on breast cancer data.

<div align="center">Table 5.1 Result Comparison with Majority Voting on Breast Cancer Data</div>

	Dempster Shafer combination misclassifications	Majority vote misclassifications
Test case1	7	7
Test case 2	10	14
Test case 3	7	9
Test case 4	5	4

Majority vote technique misclassifies 34 test records whereas Dempster's combination approach misclassifies 29 test records. Such comparison was also carried out on dermatology data. Following table shows the results on dermatology data

Table 5.2 Result Comparison with Majority Voting on Dermatology Data

	Dempster Shafer combination misclassifications	Majority vote misclassifications
Test case1	4	8
Test case 2	0	8
Test case 3	5	5
Test case 4	2	5
Test case 5	2	4
Test case 6	7	5

Majority vote technique misclassifies 35 test records whereas Dempster Shafer combination approach misclassifies 20 test records. Thus this comparison indicates that Dempster Shafer approach performs better than majority vote classifier fusion technique.

<u>Comparison with Weighted Linear Combination Technique</u>

Weighted linear combination technique is also used in classifier fusion. This techniique was applied to breast cancer data and dermatology data and results are compared with the Dempster Shafer combination approach. The weights used are classification accuracy for each class measured in past performances. Following are the accuracy measurements for breast cancer and dermatology data.

Table 5.3 Breast Cancer Data Accuracy Measurements

	kNN classifier accuracy %	Bayes classifier accuracy %	Decision tree classifier accuracy %
Class 0	96.7	94.0	95.0
Class 1	83.0	96.0	84.5

Table 5.4 Dermatology Data Accuracy Measurements

	kNN classifier accuracy %	Bayes classifier accuracy %	Decision tree classifier accuracy %
Class 0	62.0	100	98.5
Class 1	15.0	87.7	91.1
Class 2	98.0	100	99.0
Class 3	0	83.3	81.9

Class 4	0	95.8	98.5
Class 5	16.0	93.3	93.3

Classifiers are provided with same training and test data files. Each classifier proviels its belief to each class. These classes are then combined linearly using the weights. Accuracy of a classifier for each class forms the weight. The class with maximum combined belief forms the final classification.

Consider an example of a dataset with two classes. Combined beliefs of classes are evaluated as follows [20]

$$belief[0] = kNN\_weight[0] \times kNN\_belief[0] + bayes\_weight[0] \times bayes\_belief[0] + dtree\_weight[0] \times dtree\_belief[0]$$

$$belief[1] = kNN\_weight[1] \times kNN\_belief[1] + bayes\_weight[1] \times bayes\_belief[1] + dtree\_weight[1] \times dtree\_belief[1]$$

Above-mentioned weighted linear combination was applied to the breast cancer data set and dermatology dataset. Following tables provide the comparison results

Table 5.5 Results of Comparison with Weighted Linear Combination on Breast Cancer Data

	Dempster Shafer combination misclassifications	Weighted linear combination misclassifications

Test case 1	7	8
Test csse 2	10	14
Test case 3	7	6
Test case 4	5	6

Thus overall misclassifications of Dempster's approch are 29 whereas those for weighted linear combination approach are 34.

Table 5.6 Results of Comparison with Weighted Linear Combination

on Breast Cancer Data

	Dempster Shafer combination misclassifications	Weighted linear combination misclassifications
Test Case 1	4	8

Test Case 2	0	8
Test Case 3	5	5
Test Case 4	2	3
Test Case 5	2	2
Test Case 6	7	4

Thus overall misclassifications of Dempster Shafer approach are 20 whereas those of weighted linear combination are 30. This indicates that Dempster Shafer combination approach shows better performance than weighted linear combination approch.

5.3 Related Work

Different classification techniques are used in the classification of medical data. Examples are wavelets, fractal theory, artificial neural network, fuzzy theory, association rules etc. A system proposed by Guvenir and Emeksiz [ref] uses Nearest neighbor, Bayesian classifier and voting feature interval for differential diagnosis of erythmato squamous disease. It provides final diagnosis and explanation from each classifier to the doctors and students. For the detection of breast cancer tumors neural networks and association mining technique was used by Antonie, Zaiane, and Coman [ref] . All these methods provide significant performance results.

The approach introduced in this research work makes use of these individual classifiers. It fuses the results and tries to improve upon the results of the individual approaches. Some classifiers work better for a particular datasets whereas show a poor performance for some other data sets. In such cases relying on a single classifier may

lead to misclassifications as the datasets are changed. Dempster Shafer approch improves upon this inability of a classifier to work on multiple datasets. The classifiers, which can represent their classification results in terms of the beliefs to different classes, can be used as individual classifiers and fused with the other classifiers providing improved accuracy. Advantage of Dempster Shafer combination approach is that it also takes into consideration uncertainty. Combination classifier classifies record as uncertain if there is a conflict in individual classifier results. Uncertain classification doesn't qualify as misclassification but demands for more detailed medical investigation.

REFERENCES

[1] Jonathan C. Prather, David F. Lobach, Linda K. Goodwin, Joseph W. Hales ,
Marvin L. Hage¸ W. Edward Hammond Medical Data Mining: Knowledge
Discovery in a Clinical Data Warehouse 1997.

[2] Haralick, R. and Shapiro, L. (1993b).
Computer and Robot Vision, volume 2.
Addison-Wesley.

[3] Liu Rujie, Yuan Baozong, "A D-S Based Multi-Channel Information Fusion
Method Using Classifier's Uncertainty Measurement" Proceedings Of ICSP2000,
pp. 1297-1300.

[4] James C. Hoffman and Robin R. Murphy "Comparison of Bayesian and Dempster-
Shafer Theory for Sensing: A Practitioner's Approach" SPIE Proceedings on Neural
and Stochastic Methods in Image and Signal Processing II, July 1993.

[5] Horace H. S. Ip and James M. C. Ng, "Human Face Recognition Using Dempster-
Shafer Theory" IEEE 1994

[6] Robin R. Murphy, "Dempster-Shafer Theory for Sensor Fusion in Autonomous
Mobile Robots" IEEE, 1998

[7] Michel Chapron, "A Color Edge Detector Based on Dempster-Shafer Theory",
IEEE 2000

[8] Qiang Ji, Michael M.Marefat and Paul J. Lever, "An Evidential Approach for Recognizing Shape Features", IEEE 1995

[9] H. A. Güvenir and A. Akkus, Weighted K Nearest Neighbor Classification on Feature Projections in: Proceedings of the Twelfth International Symposium on Computer and Information Sciences (ISCIS XII) S. Kuru, M.U. Caglayan and H.L. Akin (Eds.), Antalya, Turkey, (Oct. 27-29, 1997), 44-51.

[10] Christian Borgelt Working Group Neural Networks and Fuzzy Systems
     Department of Knowledge Processing and Language Engineering
     School of Computer Science
     Otto-von-Guericke-University of Magdeburg

[11] http://www.plant-maintenance.com/articles/Dempster-Shafer.shtml

[12] O. L. Mangasarian and W. H. Wolberg: "Cancer diagnosis via linear programming", SIAM News, Volume 23, Number 5, September 1990, pp 1 & 18.

[13] William H. Wolberg and O.L. Mangasarian: "Multisurface method of pattern separation for medical diagnosis applied to breast cytology", Proceedings of the National Academy of Sciences, U.S.A., Volume 87, December 1990, pp 9193

[14] Nilsel Ilter, M.D., Ph.D.,
     Gazi University,
     School of Medicine, Ankara, Turkey

[15] H. Altay Guvenir, PhD.,
     Bilkent University,
     Department of Computer Engineering and Information Science, Turkey

[16] "Mathematical Techniques in Multisensor Data Fusion" David Hall Artech House Boston, London 1992

[17] A Mathematical Theory of Evidence by Glenn Shafer, Princeton University Press.

[18] Color Atlas and Synopsis of Clinical Dermatology by Fitzpatrick, Wolff, Suurmond, McGraw-Hill 4th Edition 2001.

[19] T. Denoeux. "A k-nearest neighbor classification rule based on Dempster-Shafer theory". IEEE Transactions on Systems, Man and Cybernetics, 25(05):804-813,1995.

[20] Hakan Altiçay, Mübeccel Demirekler "Comparison of different objective functions for optimal linear combination of classifiers for speaker identification". IEEE 2001

[21] http://www.ics.uci.edu/~mlearn/MLRepository.html

[22] http://www.bcff.org/Statistics.htm

[23] http://www.dermoncology.com/general/stats.html