Classification with Hidden Markov Model

Badreddine Benyacoub

Laboratory of mathematics, computer science & applications
University Mohammed V-Agdal, Faculty of Sciences
Rabat, BP 1014, Morocco

Souad ElBernoussi

Laboratory of mathematics, computer science & applications
University Mohammed V-Agdal, Faculty of Sciences
Rabat, BP 1014, Morocco

Abdelhak Zoglat

Laboratory of applied mathematics
University Mohammed V-Agdal, Faculty of Sciences
Rabat, BP 1014, Morocco

EL Moudden Ismail

Laboratory of mathematics, computer science & applications
University Mohammed V-Agdal, Faculty of Sciences
Rabat, BP 1014, Morocco

Copyright © 2014 Badreddine Benyacoub et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Classification and statistical learning by hidden markov model has achieved remarkable progress in the past decade. They have been applied in many areas like speech recognition and handwriting recognition. However, learning by Hidden Markov Model (HMM) is still restricted to supervised problems. In this paper, we propose a new learning method
based on HMM techniques estimations, to build a model for classification. The approach consists of evaluation of the probability to belonging in one group, given the observations by a linear classifier. Our developed algorithm is based on discrete states and discrete observations cases of HMM. Experimental results show that the new method has strong performance.

**Keywords:** Classification, Hidden Markov Model, Supervised learning

## 1 Introduction

Classification, one of the popular learning problems, has been applied in many areas. Many methods and algorithms have been developed for supervised learning problems, including statistical models, decision and regression trees, rules, connectionist networks, probabilistic networks. Supervised classification forms the core of what we have recently come to call data mining. The methods originated in statistics in the early nineteenth century, under the moniker discriminant analysis. An increase in the number and size of databases in the late twentieth century has inspired a growing desire to extract knowledge from data, which has contributed to a recent burst of research on new methods, especially on algorithms. Many techniques and algorithms was developed to construct linear or non-linear classifiers. There are now a multitude of ways to build classification rules, each with some common elements. The applications of supervised classification in real life is very large, like Face detection, Signature recognition, Customer discovery, spam detection, Systems Biology etc. An expert labels a set of data (training-set) in a limited number of groups (classes or labels). Labeled data is used to learn a model (classifier) in order to classify new unseen data in the defined groups. Recently, classification was an important task of data mining and machine learning, and many algorithms was developed to construct linear or non-linear classifiers [1],[8]. All study are aims to present a good predictive model.

In this study, we propose a new way of computational approach to learning, by investigating the estimation techniques of HMM. Let \((\Omega, F, P)\) be a probability space on which all the random variables to be encountered below are defined. In order, to find an efficient model, we use the basic observation equation for a Hidden Markov Model : \(Y_t = CX_t\), to develop a learning method can be handle the information present in labeled data \(Y_t\) to estimate the unseen variable \(X_t\). Our strategy works robustly in finding the satisfactory optimal estimation algorithm under the discrete HMMs constraints, which are discrete in the state, and in the measurement space. HMM satisfy different factorization and splitting properties of conditional probabilities of the bivariate process \((X_t, Y_t)\). These properties are related to the different concepts of
a stochastic system, and different constuctions based on recursive filtering and prediction approaches are proposed to solve problems of finite stochastic systems [14], [15]. Here, we discuss a finite model of hidden Markov chains where the markovian behaviour is not included in the proposed classification algorithm. Consequently some stochastic processes properties are not applied in this paper.

HMM models are also used in classification problems such as Customer credit scoring [6], Credit Risk Analysis [7], [13]. It appears that the conditional probability $P(x|y)$ is not modeled by HMM as a linear discriminant function. However, in this work, we develop a model for classification and decision by estimating the probability to belonging to one group. We use HMM as a non parametric estimator to obtain a specific model for the marginal probability $P(x)$. We determinate a set of linear functions corresponding to each group, and identifying the expression of estimate probability. Our method is available for not only classification problems but also regression problems as well as statistical models.

After this introductory part, we present the new method in Section 2, including its learning objective, probabilistic model, use of generalized inverse to construct model, and building classification model. In Section 3, we point out the connections and differences between our method and other conventional techniques related work. Experimental settings and results are given in Section 4. Finally we conclude the paper in Section 5.

2 Learning by HMM

Our goal is to estimate, from observed data, the probable group of each member $i$, where $i = 1, 2, \ldots, n$ that can be affected. The probability to be in one group $g_j$, where $j = 1, 2, \ldots, N$ is presented by a chain $\{X_i\}_{1 \leq i \leq n}$ with state space $S = \{g_1, g_2, \ldots, g_N\}$; where each state represent a group category. The process defined bellows is estimated by an observation vector $Y = \{Y(k), k = 1, \ldots, p\}$, where $Y(k) \in \{y_{1k}, y_{2k}, \ldots, y_{m(k)k}\}$, for any $1 \leq k \leq p$. In probabilistic classification analysis, we want to find $P(g_j|i)$, the probability of assigning the ith sample to jth group, where $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, N$.

2.1 Learning objective

We suppose that the sequence $X_i, 1 \leq i \leq n$ can be described by a chain, but we do not observe this chain directly. Another important supposition is that the sequence $\{X_i\}_{1 \leq i \leq n}$ is not considered as a markov chain, because either element from data has a particular information independent to the other observations. So, we suppose that $X_i$ is independent via the presence of other observations.
Denote the probability to be in one group by $\pi_j, 1 \leq j \leq N$. Each member from the data set of our study is labeled by $X_i, 1 \leq i \leq n$. So we have: $P(X_i = g_j) = \pi_{ij}$. In the following, we present a new schemes of estimating probability. We use the techniques estimation of hidden markov model to modulate those probabilities $\pi_{ij}$, using the data presented by $p$ characteristics.

Some of our work was inspired by the Hidden Markov Models: Estimation and control [2] which was used in the discrete-time and discrete-state HMM. Fundamental techniques and notation employed throughout this paper are described in [2]. Our techniques estimation is suitable to this approach.

In order to present our model, we identify the elements of $S$ with the standard unit vectors $\{g_1, g_2, \ldots, g_N\}$, where $g_j = (0, \ldots, 0, 1, 0, \ldots, 0)^t \in \mathbb{R}^N$, (t indicate the transpose), and each element $X_i$ from the process is defined by a vector of indicator functions, with $X_i^j = \langle X_i, g_j \rangle$. Note that from the definition of expectations of a simple random variable, as in [2]:

$$P(X_i = g_j) = E[\langle X_i, g_j \rangle] \quad (1)$$

For the same, each characteristic $Y(k), 1 \leq k \leq p$ can be expressed as a vector of indicator functions indicate his categories. Thus, we have

$$P(Y(k) = y_{jk}) = E[\langle Y(k), y_{jk} \rangle] \quad (2)$$

where $y_{jk} = (0, \ldots, 0, 1, 0, \ldots, 0)^t \in \mathbb{R}^{m(k)}$

Define a new matrix $C(k) = (c_{ji}(k))$, where

$$c_{ji}(k) = P(Y(k) = y_{jk} | X_i = g_i), \quad 1 \leq i \leq N, \quad 1 \leq j \leq m(k) \quad (3)$$

$C(k)$ is the matrix of transition probabilities whose entries satisfy $\sum_{j=1}^{m(k)} c_{ji}(k) = 1, c_{ji}(k) \geq 0$. Then

$$E[Y(k) | X_i] = C(k)X_i \quad (4)$$

If we define $W(k) = y(k) - C(k)X_i$, we have $E[W(k) | X_i] = 0 \in \mathbb{R}^{m(k)}$, where $W(k)$ is a noise process due to failure of estimation. The various of $Y$ can then be written as

$$Y(k) = C(k)X_i + W(k) \quad (5)$$

The learning target is now to find an estimation of the matrix $C(k)$. The following equation presents a method to obtain estimates of parameters as indicated in [12]

$$\hat{c}_{ji}(k) = \frac{\text{expected number times in state } i \text{ and observing symbol } y_j(k)}{\text{expected number times in state } i} \quad (6)$$

The estimator $\hat{C}(k)$ is called the maximum likelihood estimate of HMM.

In other word, we suppose that the observed process contains information about the assigned group of each element of data.
2.2 Probabilistic model

Our model for the process \( X_i, 1 \leq i \leq n \) is as follows:

**Hidden Markov Model**

\[
Y(k) = C(k)X_i + W(k)
\]

\( C(k) \) is the matrix of transition probabilities, who satisfy

\[
\sum_{j=1}^{m(k)} c_{ji}(k) = 1, \quad \text{and} \quad c_{ji}(k) \geq 0.
\]

Our processes are defined on a probability space \((\Omega, F, P)\).

**Observation equation**

Now we represent the above equations by a vector as following

\[
C = \begin{pmatrix}
C(1) \\
C(2) \\
\vdots \\
C(p)
\end{pmatrix}, \quad Y = \begin{pmatrix}
Y(1) \\
Y(2) \\
\vdots \\
Y(p)
\end{pmatrix}
\]

and

\[
W = \begin{pmatrix}
W(1) \\
W(2) \\
\vdots \\
W(p)
\end{pmatrix}
\]

Therefore we can write:

\[
Y = CX_i + W \quad (7)
\]

Let \( m = m(1) + m(2) + \ldots + m(p) \), then \( C \) is a \( m \times N \) matrix. Multiple this equation by transpose matrix \( C^t \)

\[
C^tY = C^tCX_i + C^tW \quad (8)
\]

Write \( M = C^tC \), \( M \) is a \( N \times N \) symmetric matrix.

If \( C \) is a full rank matrix (all columns are linearly independent), we can calculate the matrix inverse \( M^{-1} \).

let

\[
M^{-1}C^tY = X_i + M^{-1}C^tW \quad (9)
\]

Finally, we obtain:

\[
X_i = M^{-1}C^tY - M^{-1}C^tW \quad (10)
\]

The matrix \( A = M^{-1}C^t \) gives the model parameters, and the matrix \( M^{-1}C^tW \) represent the margin error due for estimation.
So, we can write:

\[ <X_i, g_i> = \sum_{k=1}^{p} \sum_{j=1}^{m(k)} a_{jk}^i <Y(k), y_{jk}>, \quad i = 1, \ldots, N \] (11)

We are developing a linear model for classification where the matrix \( M = C^t C \) may be non-invertible. In this context, the generalized inverse provides a solution to this problem as shown in the next section.

### 2.3 Use of Generalized Inverse to Construct Model

The purpose of constructing a generalized inverse is to obtain a matrix that can serve as the inverse in some sense for a wider class of matrices than invertible ones. The importance of the generalized inverse matrix \( G \) is revealed in the following definition and theorem.

#### 2.3.1 Definition and Properties of Generalized Inverse

**Definition** A generalized inverse of a matrix \( A \) of order \( m \times n \) is a matrix of order \( n \times m \), denoted by \( A^+ \), such that for any \( y \) for which \( Ax = y \) is consistent, \( x = A^+y \) is a solution.

An inverse so defined is not necessarily unique, but may be made unique by imposing further conditions. Some of the earlier writers considered the problem of defining a unique inverse and studying its properties.

**Theorem** [10]: \( A^t \) is a generalized inverse of \( A \) then \( AA^t A = A \) and conversely.

Choose \( y \) as the \( i \)th column \( a_i \) of \( A \). Then the equation \( Ax = a_i \) is obviously consistent and hence \( x = A^t a_i \) is a solution, i.e \( AA^t a_i = a_i \) for all \( i \), which implies that \( AA^t A = A \).

Conversely if \( A^+ \) exists such that \( AA^t A = A \), and \( Ax = y \) is consistent, then \( AA^t Ax = Ax \) or \( AA^t y = y \). Hence \( x = A^t y \) is a solution.

Although \( A^t \) has infinitely many variety satisfy this theorem, any one of them will do for our purposes. However, Moore and (unaware of Moor’s work) [12] reduced the infinity of generalized inverses to one unique solution by imposing four reasonable algebraic constraints, all met by the standard inverse. If

1. general condition: \( AA^t A = A \),
2. reflexive condition: \( A^t AA^t = G \),
3. normalized condition: \( (AA^t)^t = AA^t \), and
4. reverse normalized condition: \( (A^t A) = AA^t \)
then this $A^t$ matrix is unique. The proof is lengthy, and we refer the interested reader to Penrose [12].

Because the properties of the Moore-Penrose generalized inverse are intuitively desirable, and because of the invariance of the important statistical results to the choice of generalized inverse, we follow standard statistical practice by using this form from now.

### 2.3.2 Application to our problem

Our model for classification is presented by observation equation as follows:

$$Y = CX_i + W$$

Where $E[X_i] = \pi_i = (\pi_1, \pi_2, \ldots, \pi_N)^t$. Therefore, we have a system of $m$ linear equation such that

$$E[Y] = C\pi_i \quad (12)$$

Where $C$ is $m \times N$ matrix with estimated coefficients from a matrix data of $n$ observations and $\pi_i$ is $N \times 1$ vector of unknown parameters. Let $G$ be any g-inverse of $C^tC$. Then $t = GC^tE[Y]$ is a system of equations which can estimate $E[X_i]$. If $C^tC$ is not invertible, the solution is not unique, but, however, a linear function $p^t\pi$ may be unique. Such a function is said to be estimable and its estimate is given by $p^t\pi = p^tGC^tE[Y]$. We have the following result.

$$E[p^t\pi] = p^t\pi \quad (13)$$

if $p^t\pi$ is estimable.

Since $p^t\pi$ is unique we have, by lemma 2, $p^tGC^tC = p^t$. Thus

$$E[p^t\pi] = E[p^tGC^tE[Y]] = p^tGC^tC\pi = p^t\pi$$

which proves the above relation. This result show that $t$ may be regarded as estimate of $\pi$, for purposes of building up an estimate of any estimable function $p^t\pi$.

Now, let $M^+$ the generalized inverse of the matrix $M = C^tC$. The implementation of generalized inverse to build our model yield this equation:

$$X_i = M^+C^tY - M^+C^tW \quad (14)$$

The resulting equation produces classification rules as the same presented in Both the fitted values of $Y$ and the residual error variance are invariant to the choice of $M^+$.

### 2.4 Building classification model

HMM are used in this work to define a new way of classification and make-decision. In the following, we present the process to build our model, and the
completed through several steps to evaluate and classify an element. Taking the expectation of the expression (10) as mentioned in section (2.2). Then we have N equations, which define the relation between estimated probabilities \( \{\pi_i, i = 1, 2, \ldots, N\} \) and the observed characteristics \( \{Y(k), k = 1, 2, \ldots, p\} \).

The probability to belonging in one group is determined by corresponding equation. We use the associated coefficients corresponding to the categories that verify for each equation to estimate the the probability.

Now observe that taking the expectation of each equation presented in expression (10), we have

\[
E[M^{-1}C^tW] = M^{-1}C^tE[W] = 0.
\]

Consequently, we assume that, the noise terms \( W \) in the model does not involves in classification procedure. Finally, We obtain those N following probabilities estimation.

\[
\begin{align*}
P(X = g_1) &= a_{11}^1P(Y(1) = y_{11}) + \ldots + a_{jk}^1P(Y(k) = y_{jk}) \\
&\quad + \ldots + a_{M(p)M}^1P(Y(p) = y_{M(p)p}) \\
P(X = g_2) &= a_{11}^2P(Y(1) = y_{11}) + \ldots + a_{jk}^2P(Y(k) = y_{jk}) \\
&\quad + \ldots + a_{M(p)M}^2P(Y(p) = y_{M(p)p}) \\
&\vdots \\
P(X = g_N) &= a_{11}^N P(Y(1) = y_{11}) + \ldots + a_{jk}^NP(Y(k) = y_{jk}) \\
&\quad + \ldots + a_{M(p)M}^NP(Y(p) = y_{M(p)p})
\end{align*}
\]

(15)

A supervised classification algorithm has a process, including training model, testing model. This algorithm can be widely used in two steps, and on two different parts of dataset. Details of these steps are be presented in the following.

The training algorithm has as aim to learn the matrix \( C \), from a selected sample using the dataset. Once the coefficients of \( C \) are estimated, then we can build our model implementing the procedures followed in the HMM method of calculation as shown in the section (2.2). As results, we get a discrimination equation corresponding to each group. In implementation steps, we first convert continuous data to categorical data. Discretization of continuous attributes is an important task for our process. Many discretization algorithms such as Boolean reasoning algorithm, entropy algorithm and naive algorithm have been proposed to deal with this problem. In order to efficiently obtain those discriminants functions, the data set is preprocessed by discretization. The input data is organized as a matrix to be suitable for our algorithm, where each column represent an observed characteristic. Partitions ranges for each continuous attributes can be determined by applying a discretization algorithm on the values generated in training data. For the purposes of the algorithm, each ranges in the output is represented in its categorical form.

In testing algorithm, we use the remainder dataset to test classification accuracy of model. At this stage, we use the above equations to classify a new observation from test data. Each equation is expressed by a set of probabilities and associated coefficients. Noting that \( P(Y(k) = y_{ik}) \) takes the values 0 or 1 to indicate the absence or presence of each categorical in the model. The
procedure of classification consist to assign each element from test data to the probable group. This procedure algorithm has the following steps: 1) A collection of information for an observation is known as a set of categories, 2) A selection from classification equations of associated coefficients corresponding to the set of categories, 3) A summation of those associated coefficients for each equation gives the probabilities to belonging in each group, 4) A identification of the probable group by comparing the estimated probabilities taking the greatest of them.

3 Related Work

Our method intersects with several other machine learning approaches. Here we discuss some of these directions, pinpointing the connections and our new contributions.

3.0.1 Linear discriminant analysis (LDA)

LDA method is a statistical discriminative way. Let the probability that an element with measurement vector $y$ belongs to group $g$ is $P(g_i|y)$. We consider the set of distributions $\{P(g_1|y), P(g_2|y), \ldots, P(g_N|y)\}$ which are multivariate normal distributions $N(\mu_i, \Sigma_i)$ with the common variance $\Sigma_1 = \Sigma_2 = \ldots = \Sigma_N = \Sigma$. Each element was classified into the group $g_i$ where the posterior probability $P(g_i|y)$ that it was a member of that group, given its value of $y$, was largest. LDA should apply Bayes formula to evaluate $P(g_i|y)$ using $P(y|g_i)$ and the prior probability $P(g_i)$. This posterior probability was calculated using Baye’s rule as

$$P(g_i|y) = \frac{P(y|g_i)p_i}{\sum_{k=0}^{N} P(y|g_k)p_k} \quad (17)$$

where $p_k$ is the prior probability that a case is a member of group $k$.

3.0.2 Logistic regression (LR)

logistic regression is used under the assumption that the posterior probability that an element is a member of group $k$ is specified directly as

$$P(g_i|y) = \frac{(\exp(\beta_k^Ty))}{\sum_{j=0}^{r} \exp(\beta_j^Ty)} \quad (k = 0, 1, \ldots, r) \quad (18)$$

where $\beta_k$ is a column vector of coefficients for group $k$. To make the parameter estimates identifiable, it is usual to normalize the coefficients for one group. The model of LR so far described is often known as the multinomial logit model. The logarithm of the likelihood function can be formulated and
differentiated to give estimators of the parameters. In this work the parameter vectors were estimated using the method of maximum likelihood, and a case was classified into the group of which it had the highest probability of membership. The detailed settings of the compared methods are as follows.

### 3.1 Discussions and comparison of three methods

We have compared our method with two other classification method that can take a numerical matrix data of as input. The compared algorithms range from classical to state of the art method with various principles: LDA uses Baye’s rule to predict the posterior probability. Unlike LDA, LR inspires the multinomial logit model to model the posterior probability, and uses the above expression as posterior probability. In contrast, our method directly learns the assigning posterior probability \( P(g_i|y) \). Our HMM-model, based on estimation of matrix C use the likelihood function to calculate the parameters of matrix C. The expression (6) is a result of estimation procedure by maximum likelihood as mentioned in [12].

Eisenbeis Avery [5] and Eisenbeis [3],[4] discuss eight problems with using LDA in classification supervised problems. For example, the linear discriminant model assumes (a) that the discriminating variables are measured on an interval scale, (b) that the covariance matrices of the discriminating variables are equal for the groups, and (c) that the discriminating variables follow a multivariate normal distribution. Several datasets used in supervised learning problems contain discriminating variables of nominal order and so assumptions (a) and (c) are violated. It is well known that, when predictor variables are a mixture of discrete and continuous variables, the linear discriminant function may not be optimal. The logistic regression model and HMM-model developed in this work, does not require the assumptions necessary for the linear discriminant model. So the two methods are suitable for any form of datasets. A major advantage in HMM-model, unlike supervised learning techniques does not need to provide inference for data in classification analysis.

### 4 Experiments

#### 4.1 Datasets

The performance of classification methods were evaluated using real-world datasets. In particular, we focus on data that contain variety of characteristics continuous, discrete and categorical data. We thus selected 4 such datasets which are publicly available from a variety of domains.

The datasets are composed as a matrix, such that each row correspond to one observation, and each column correspond to one attribute as an input
feature. A this stage, every line of datasets stand for a case of membership to one group. Each line can be considered as a row observation vector \( y = (y_1, y_2, \ldots, y_p) \), where represent \( p \) measured variables of dataset. According to HMMs context, this row will be the sequence of observed data or input symbols, and each element being emitted by a hidden state represented here by \( \{g_1, g_2, \ldots, g_N\} \). The set of state is varied from one dataset to another.

The \( p \) measured variables are continuous attributes, discrete attributes, categorical attributes, or a mixture of them. The datasets are re-organized such that all observed variables are in discrete categorial values. Therefore the data are converted to a suitable form for HMM. Implementation of our model start from a data processing step. It was an intermediate step before building the model. At this stage, each observed characteristics is quantified to a specific number of categories. Suppose, that \( y_1 \) contains \( m_1 \) classes, \( y_2 \) contains \( m_2 \) classes, and so on. The number of categories represent distinct observation symbols per state of the HMM model was the sum of all classes such that: \( M = \sum_{i=1}^{p} m_i \).

### 4.2 Discussion of results and Analysis

Classification performance for three methods, namely HMM, LR, LDA, will be compared by accuracy rate. Accuracy is the most direct criteria, to evaluate classification models. It can be measured quantitatively, as following:

\[
\text{Accuracy} = \frac{\text{The number correctly classified cases}}{\text{The total number of cases}} \tag{19}
\]

The experiments are performed with training data and test data, where the size is chosen differently and depend to available dataset. To provide a reliable estimate and validate the developing models, tenfold cross-validation was adopted. The entire dataset was randomly split into subsets, one for training and other for testing. Each time, the model was built by first subset, as the training sample and the remaining sample was retained for testing. Finally, the average of the 10 results was taken as the prediction accuracy. Classification accuracy will be calculated for 10 test data. The results determined for each data are shown in the table 1.

<table>
<thead>
<tr>
<th>DATASET</th>
<th>HMM</th>
<th>LDA</th>
<th>LR</th>
</tr>
</thead>
<tbody>
<tr>
<td>IRIS</td>
<td>95.60</td>
<td>97.20</td>
<td>96.80</td>
</tr>
<tr>
<td>VOTES</td>
<td>97.50</td>
<td>94.40</td>
<td>90.00</td>
</tr>
<tr>
<td>GERMAN</td>
<td>76.10</td>
<td>72.02</td>
<td>75.12</td>
</tr>
<tr>
<td>AUSTRALIAN</td>
<td>86.60</td>
<td>86.37</td>
<td>86.52</td>
</tr>
</tbody>
</table>

Table 1: Classification Accuracy by HMM, LDA and LR.
We use MatLab to compare the models used in our experiments. Logistic regression (LR) was performed with MatLab function "mnrfit". This function returns a matrix of coefficient estimates, which represents parameters of model. Then, we use returned matrix in MatLab function ”mnrval” to compute predicted probabilities. For linear discriminant analysis (LDA), ”classify” MatLab function is employed to classify each row of the data. The output of this function is a vector indicates the group to which each row of sample has been assigned. For HMM-model, the input parameters include the number of states and the number of observation symbols. At this stage, the states number N is equal to the number of classes defined in each dataset. The observation numbers M was determined by the number of classes in the attributes used. The two parameters N and M determine the size of the learned matrix C.

The results shown in Table.1 are an average of 10 repetitions. The results show that, the best accuracy is obtained by LDA 97.20% for Iris dataset, which is followed by 96.80% and HMM 95.60%. For Votes, German and Australian datasets, the highest accuracy is obtained from developed HMM model, than LR and LDA with values 97.50%, 76.10% and 86.60% respectively. LR follows HMM in terms of classification accuracy for two datasets German and Australian with values 75.12% and 86.52% respectively.

Table 1. presents the experimental results of the three models; HMM, LR and LDA over the four datasets. Note from this table that the highest average accuracy is obtained by HMM for Votes, German and Australian datasets. However, LDA has the best accuracy for Iris dataset. It can be concluded that the HMM model yields a very good performance with datasets contain categorical attributes, unlike LDA, which is an efficient method when all attributes are continuous.

5 Conclusion

We have proposed a hidden Markov model for classification, and we make it a method for decision as well as traditional methods. Our method has three major original contributions: (1) a novel probabilistic approach for building discriminant functions derived from discrete HMMs (Discrete States and Discrete observations; (2) a strategy uses a selected method of categories for finding probabilities; (3) a learning analysis could be beneficial to construct a classification model without including statistical assumptions in implementing discriminant analysis. The experiment of this model is proved to its simplicity in implementation and evaluated its significant value en terms of classification. Our method works robustly for selected various datesets and can improve classification for large manifolds datasets.
References


Received: February 25, 2014