2010 International Conference on Pattern Recognition

Maximum entropy model based Classification with Feature Selection

Ambedkar Dukkipati, Abhay Kumar Yadav and M. Narasimha Murty
Dept. of Computer Science and Automation
Indian Institute of Science
Bangalore 560012, India
ambedkar@csa.iisc.ernet.in

***Abstract*—In this paper, we propose a classification algorithm based on the maximum entropy principle. This algorithm finds the most appropriate class-conditional maximum entropy distributions for classification. No prior knowledge about the form of density function for estimating the class conditional density is assumed except that the information is given in the form of expected valued of features. This algorithm also incorporates a method to select relevant features for classification. The proposed algorithm is suitable for large data-sets and is demonstrated by simulation results on some real world benchmark data-sets.**

***Keywords*-Bayes; Jefferys divergence; sample mean;**

I. Introduction

Classification problem can be stated as follows: given a set of N training data points $(x_i,\ y_i)$, $i = 1, \ldots, N$, $x_i \in X$, $y_i \in Y$, the goal is to find the underlying unknown mapping or decision function $h : X \rightarrow Y$. Most Commonly used classifiers are support vector machines (SVM) [1], K-nearest neighbors [2], Bayes classifier [3], Adaboost [4], decision trees, neural networks (multi-layer perceptron) and so on.

The classification algorithms can be broadly classified as linear classifiers and nonlinear classifiers. The decision function or decision boundary of a linear classifier is the linear combination of the features. The most common examples of linear classifier are support vector machines (SVM) [1], Fisher's linear discriminant [5] and a single layer perceptron [6].

A classifier is said to be a nonlinear classifier, if its decision surface is a nonlinear combination of the features, for example, quadratic classifier. A quadratic classifier can be a decision function,

$$f(\mathbf{x}) = \mathbf{x^T A x} + \mathbf{b^T x} + c$$

where, $\mathbf{A} \in R^d \times R^d, b \in R^d$ and $c \in R$, such that if $f(\mathbf{x}) > 0$, the class label assigned is c_1, otherwise c_2. Bayes classifier[3] is a quadratic classifier when the multivariate normal distribution is used as the probabilistic model for both the classes.

Bayes classifier is specified in terms of the class conditional densities. Once the class conditional densities are estimated (parametric or non-parametric), Bayes classifier assigns a class label to a test pattern/observation as follows: Let $P(c_1)$ and $P(c_2)$ be the prior probabilities of the two classes and $P(\mathbf{x}|c_1)$ and $P(\mathbf{x}|c_2)$ be the class conditional densities for class c_1 and c_2 respectively. Then Bayes classifier assigns test pattern to the class c_1 if,

$$P(\mathbf{x}|c_1)P(c_1) > P(\mathbf{x}|c_2)P(c_2) \tag{1}$$

otherwise assign test pattern to the class c_2. Also in this case two class classification problem can easily be extended to multiclass classification problem.

Estimation of class conditional density mainly includes parametric and non-parametric approaches. In a parametric approach, some form of class conditional densities are assumed *a priori*.

The most commonly used model is Gaussian model. But sometimes data may not fit well to the Gaussian model. In this respect, maximum entropy principle offers more general and flexible models.

Maximum entropy (ME) principle has been used to learn statistical models in many applications such as natural language processing [7], texture modeling [8]. In this paper, we propose a method based on ME principle to estimate the class conditional densities.

The paper is organized as follows. Section II gives the information theory background. We present our proposed method in Section III. In Section IV, we present simulation results on some real and artificially generated datasets.

II. Maximum Entropy Fundamentals

The maximum entropy principle (Jaynes, 1957) states that we should make use of all the information that is given and scrupulously avoid making assumptions about information that is not available. Let $\mathbf{X}$ be a random vector *i.e.*, $\mathbf{X} = (x_1, \ldots, x_d),\ x_i \in \chi_i,\ i = 1, \ldots, d$. Aim is to estimate the distribution using the information given in the form of expected values of some moment functions $C = \{\phi_1(\mathbf{x}), \ldots, \phi_m(\mathbf{x})\}$. According to ME principle, out of all those distributions consistent with the given constraints, we should choose the distribution that maximizes Shannon entropy,
That is, we maximize

$$H = -\int P(\mathbf{x}) \ln P(\mathbf{x})\ d\mathbf{x} \tag{2}$$

1051-4651/10 $26.00 © 2010 IEEE
DOI 10.1109/ICPR.2010.143

subject to

$$\int \phi_r(\mathbf{x})P(\mathbf{x})\ d\mathbf{x} = E_P\left[\phi_r(\mathbf{x})\right],\ \ r = 1, 2, \ldots, m\ \ , \quad (3)$$

and $\int P(\mathbf{x})d\mathbf{x} = 1, \quad P(\mathbf{x}) \geq 0$. The classical solution for this problem is given by[1]

$$P(\mathbf{x}) = \exp\left(-\lambda_0 - \sum_{j=1}^{m} \lambda_j \phi_j(\mathbf{x})\right)\ \ , \quad (4)$$

where $\lambda_0, \lambda_1, ..., \lambda_m$ are obtained by solving the following set of $(m+1)$ nonlinear equations

$$\int \exp\left(-\lambda_0 - \sum_{j=1}^{m} \lambda_j \phi_j(\mathbf{x})\right)\ d\mathbf{x} = 1\ \ , \quad (5)$$

and

$$\int \phi_r(\mathbf{x}) \exp\left(-\lambda_0 - \sum_{j=1}^{m} \lambda_j \phi_j(\mathbf{x})\right)\ d\mathbf{x} = E_P\left[\phi_r(\mathbf{x})\right], \quad r = 1, 2, \ldots, m\ \ . \quad (6)$$

The expected values of the moment constraint functions *i.e.*, $E_f\phi_r(x)$ can be approximated by observed statistics or sample means of $\phi_r(\mathbf{x})$. *i.e.*,

$$E_P\left[\phi_r(\mathbf{x})\right] = \frac{1}{N}\sum_{i=1}^{N} \phi_r(\mathbf{x}_i) = \mu_r^{emp},\ \ r = 1, 2, \ldots, m\ \ . \quad (7)$$

The maximum value of entropy is given by

$$H_{max} = \lambda_0 + \sum_{j=1}^{m} \lambda_j \mu_j^{emp}\ \ . \quad (8)$$

Since entropy is the measure of uncertainty, a maximum entropy model can be thought of as a simplest model that is consistent with the given feature constraints. To compute the ME distribution one has to solve (5) and (6), which does not have the closed analytical solution. However, the max entropy distribution parameters $\Lambda = (\lambda_0, \ldots, \lambda_m)$ can be found by maximizing the likelihood function of the parametric exponential model

$$P(\mathbf{x}; \Lambda, C) = \exp\left(-\lambda_0 - \sum_{j=1}^{m} \lambda_j \phi_j(\mathbf{x})\right)\ \ .$$

[1]The moment constraint function may be vector valued functions, in that case λ_j are vector of the same length as $\phi_r(\mathbf{x})$ and $P(\mathbf{x}) = \exp[-\lambda_0 - \sum_{j=1}^{m} \lambda_j^T \phi_j(\mathbf{x})]$

III. Proposed Method

Kullback-Leibler divergence (KL-divergence) is a non-symmetric directed divergence which measures the distance of a probability distribution from the other probability distribution. The-KL divergence from $f_1(x)$ to $f_2(\mathbf{x})$ is defined as,

$$\begin{aligned} I(f_1(\mathbf{x}), f_2(\mathbf{x})) &= \int f_1(\mathbf{x}) \ln \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} d\mathbf{x} \\ &= -H(f_1(\mathbf{x})) - E_{f_1}\left[\ln f_2(\mathbf{x})\right]\ \ . \end{aligned} \quad (9)$$

The closeness of a probability model $g(\mathbf{x})$ from the true underlying distribution can be measured using the KL divergence. Let the true distribution be $f(\mathbf{x})$, then KL divergence from $f(\mathbf{x})$ to $g(\mathbf{x})$ can be written as,

$$\begin{aligned} I(f(\mathbf{x}), g(\mathbf{x})) &= \int f(\mathbf{x})\ \ln \frac{f(\mathbf{x})}{g(\mathbf{x})}\ d\mathbf{x} \\ &= -\mathcal{H}(f) - E_f\left[\ln g\right] \end{aligned} \quad (10)$$

The KL-divergence from $f(\mathbf{x})$ to a maximum entropy model $P(\mathbf{x}; \Lambda^*, C)$ is given by,

$$I(f(\mathbf{x}),\ \ P(\mathbf{x}; \Lambda^*, C)) = H(P) - H(f) \quad (11)$$

More number of constraints means that we are using more information and thus entropy will decrease with every additional constraint or information. Thus to minimize the KL divergence one should use as many constraints as possible to minimize the model entropy. In this sense, minimum entropy principle supports the generality of the model.

Symmetrized version of KL-divergence is known as Jefferys divergence or J-divergence which can be used to measure distance between probability distributions. J-divergence between two probability distributions $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$ is defined as [9] (also see [10, Chapter 3]),

$$J\left(f_1, f_2\right) = \frac{1}{2}\left\{I(f_1,\ \ f_2) + I(f_2,\ \ f_1)\right\} \quad (12)$$

Let P_1 and P_2 be two maximum entropy models consistent with some finite moment constraints sets, then J-divergence between the two models is given by,

$$\begin{aligned} &J(P_1, P_2) \\ &\quad = \frac{1}{2}\Big(- H(P_1) - E_{P_1} \ln P_2 - H(P_2) - E_{P_2} \ln P_1\Big) \end{aligned} \quad (13)$$

When we add new information in the form of a constraint, entropy of ME model decreases. From this we can deduce that the addition of a new constraint to the existing constraints set leads to increase in $J(P_1, P_2)$ and decrease in KL-divergence. Decrease in KL-divergence indicates that the ME models are closer to their respective true unknown probability distributions.

From the above discussion, one can conclude that for discrimination between two ME models one should include as

many constraints as possible, in this sense it favors the model complexity, on the other hand maximum entropy principle favors the model simplicity. Thus it can be observed that there is a trade off between the model simplicity and the model complexity.

Consider a two class problem. Let the two classes be c_1 and c_2. $\mathbf{X}$ be a random vector *i.e.*, $\mathbf{X} = (x_1, \ldots, x_d),\ x_i \in \chi_i,\ i = 1, \ldots, d$. Let the samples of class c_1 and c_2 be represented by S_1 and S_2, respectively. Suppose $\mid S_1 \mid = N_1,\ \mid S_2 \mid = N_2$. Let the true unknown underlying class conditional densities be $f^*_{c_1}(\mathbf{x}),\ f^*_{c_2}(\mathbf{x})$ and the given set of l feature constraint functions be $\Omega = \{\varphi_i(\mathbf{x}) : i = 1, \ldots, l\}$, where $\varphi_i(.)$ can be vector valued function.

Consider two arbitrary sets of constraints having sizes M_1 and M_2 constraints, let these sets be Δ_{c_1} and Δ_{c_2} for class c_1 and c_2 respectively. The statistics of these constraints are estimated by sample means. We compute the class conditional maximum entropy distributions for class c_1 and c_2 subjected to the constraints Δ_{c_1} and Δ_{c_2} and the normal probability constraints $\int\ P_{c_1}(\mathbf{x})\ d\mathbf{x} = 1,\ P_{c_1}(\mathbf{x}) > 0$ and $\int P_{c_2}(\mathbf{x})\ d\mathbf{x} = 1,\ P_{c_2}(\mathbf{x}) > 0$. Let these class conditional ME distributions be $P_{c_1}(\mathbf{x}; \Delta_{c_1})$ and $P_{c_2}(\mathbf{x}; \Delta_{c_2})$. For different sets of M_1 and M_2 moment constraint functions we will get different class conditional ME distributions. Thus out of all different class conditional ME distributions we should choose those class conditional ME distributions $P^*_{c_1}(\mathbf{x})$ and $P^*_{c_2}(\mathbf{x})$ which have the maximum discrimination power or maximum J-divergence between them. Thus we get the following optimization problem,

$$(P^*_{c_1}, P^*_{c_2}) = \underset{(\Delta_{c_1}, \Delta_{c_2}) \in B \times B}{\arg\max} J\{P_{c_1}(\mathbf{x}; \Delta_{c_1}), P_{c_2}(\mathbf{x}; \Delta_{c_2})\}\ ,$$

where B is the power set of Ω. The other interpretation is that we want those class conditional models which are simpler (maximum entropy) and also have the discriminative power (Symmetric KL-divergence). Note that it is practically infeasible to enumerate all possible subsets of feature constraints. Hence, we propose a greedy method for the problem.

Since the maximum entropy distribution depends on the constraint functions lets say we have a set C of optimal constraints. Initially, we start without considering any moment constraint function *i.e.*, $C = \{\}$. Under the assumption that sample size is large enough for both the classes the expected value for the moment constraint functions $\phi_r(\mathbf{x})$ for class c_1 and c_2 can be approximated by their respective sample means. *i.e.*,

$$E_P\ [\phi_r(\mathbf{x}; c_1)] = \frac{1}{N_1} \sum_{\mathbf{x} \in S_1} \phi_r(\mathbf{x}_i),\ \ r = 1, 2, \ldots, |C| \tag{14}$$

$$E_P\ [\phi_r(\mathbf{x}; c_2)] = \frac{1}{N_2} \sum_{\mathbf{x} \in S_2} \phi_r(\mathbf{x}_i),\ \ r = 1, 2, \ldots, |C| \tag{15}$$

The form of $P^*_{c_1}(\mathbf{x}; C)$ and $P^*_{c_2}(\mathbf{x}; C)$ will be

$$P^*_{c_1}(\mathbf{x}; C) = \exp\left(-\lambda_0 - \sum_{j=1}^{|C|} \lambda_j \phi_j(\mathbf{x})\right)$$

$$P^*_{c_2}(\mathbf{x}; C) = \exp\left(-\nu_0 - \sum_{j=1}^{|C|} \nu_j \phi_j(\mathbf{x})\right)$$

where λ_j and ν_j are the vectors of the same length as $\phi_j(\mathbf{x})$ and $|C|$ denotes cardinality of the set C

The initial class conditional distribution $P^*_{c_1}(\mathbf{x}; C)$ and $P^*_{c_2}(\mathbf{x}; C)$ set to be uniform distributions. And for next step we add one more constraint $\phi_\tau(\mathbf{x})$ from the set Ω and compute the maximum entropy distribution. Let $C^+ = C \cup \phi_\tau(\mathbf{x})$. We again compute the class conditional max entropy distribution $P^*_{c_1}(\mathbf{x}; C)$ and $P^*_{c_2}(\mathbf{x}; C)$ with respect to the new constraints set C^+, which have following forms,

$$P^*_{c_1}(\mathbf{x}; C^+) = \exp\left(-\lambda^*_0 - \left(\sum_{j=1}^{|C|} \lambda^*_j \phi_j(\mathbf{x})\right) - \lambda^*_\tau \phi_\tau(\mathbf{x})\right)$$

$$P^*_{c_2}(\mathbf{x}; C^+) = \exp\left(-\nu^*_0 - \left(\sum_{j=1}^{|C|} \nu^*_j \phi_j(\mathbf{x})\right) - \nu^*_\tau \phi_\tau(\mathbf{x})\right)$$

In general we have $\lambda^*_j \neq \lambda_j$ and $\nu^*_j \neq \nu_j$. Now, we compute the following function for constraint $\phi_\tau(\mathbf{x})$,

$$\psi(\phi_\tau(\mathbf{x})) = J\left\{P^*_{c_2}(\mathbf{x}; C^+),\ \ P^*_{c_1}(\mathbf{x}; C^+)\right\}, \tag{16}$$

and

$$\phi_{\tau^*}(\mathbf{x}) = \underset{\tau \in \Omega \setminus C}{\arg\max}\ \ \psi(\phi_\tau(\mathbf{x})). \tag{17}$$

We add the constraint $\phi_{\tau^*}(\mathbf{x})$ to the set of optimal constraint set and update $C = C \cup \{\phi_{\tau^*}(\mathbf{x})\}$ and $\Omega = \Omega - \{\phi_{\tau^*}(\mathbf{x})\}$. We keep adding the constraints in C as long as the J-divergence is greater than some threshold value. The proposed method can be summarized as follows;

Input : Two datasets $S_1,\ S_2$ with their respective class labels and set of finite constraint functions.

Initialization : We start with uniform distribution as class conditional density $P^*_{c_1}(\mathbf{x}; C)$ and $P^*_{c_2}(\mathbf{x}; C)$ for both the classes and optimal constraint subset C is empty and remaining constraint set be Ω = set of all moment l constraints

Algorithm:

1 If Ω is empty stop else to step 2.

2 $\forall\ \tau \in \Omega$, make a new constraint set $C^+ = C \cup \tau$ and let the class conditional max entropy distributions be $P^*_{c_1}(\mathbf{x}; C^+)$ and $P^*_{c_2}(\mathbf{x}; C^+)$. We add a new constraint

Problem	*#train data*	*#test data*	*#attributes*
Breast-Cancer	200	77	9
Diabetis	468	300	8
Heart	170	100	13
Splice	1000	2175	60
Titanic	150	2051	3
German	700	300	20

Table I
PROBLEM STATISTICS

τ^* to optimal subset of constraints C and set $\Omega = \Omega - \tau^*$. Where τ^* is given by,

$$\tau^* = \underset{\tau \in \Omega \setminus C}{\arg\max}\ J\left\{P^*_{c_2}(\mathbf{x}; C^+),\ \ P^*_{c_1}(\mathbf{x}; C^+)\right\}$$

3 If relative increase in the J-divergence is less then some threshold value, stop the algorithm. Otherwise go to step 1.

IV. SIMULATION RESULTS

We performed several experiments on real world data sets chosen from UCI machine learning repository and all problem are binary class classification problems. The problem statistics is given in Table I. All the problems are binary class classification problems. We used 5 fold cross validation to find the optimal constraint functions set. Several experiments were performed on the datasets and average test error is reported in Table II. These values show the average and standard deviation of error on the 100 different sets of the same data used in experiments (in some cases only 20 were available). The results are compared with benchmark results explained in [11] and shown in Table II. For most of the datasets, the performance is very competitive with the benchmark results.

V. CONCLUSION

The novelty of proposed method is that the classifier automatically chooses the relevant feature which appropriate for the classification task. The proposed classification algorithm is suitable for large datasets as it requires the information only in the form of expected values of the constraint functions, which can be computed incrementally. We have also shown that it is better to use fewer constraint functions than all the constraint functions which is also consistent with the MDL (minimum description length) or bias-variance trade off. To extend this method for multiclass classification one approach that could to be explored is to use a mixture model which has maximum entropy components as number of classes.

Problem	*Proposed Method*	*RBF -Network*	*AdaBoost C, S opt*	*SVM*
Breast -Cancer	27.2 [+/-4.7]	27.6 [+/-4.7]	30.4 [+/-4.5]	26.0 [+/-4.7]
Diabetis	26.16 [+/-1.94]	24.3 [+/-1.9]	26.5 [+/-2.3]	23.5 [+/-1.7]
Heart	19.3 [+/-3.11]	17.6 [+/-3.3]	20.3 [+/-3.4]	16.0 [+/-3.3]
Splice	10.58 [+/-0.5]	10.0 [+/-1.0]	10.1 [+/-.5]	10.9 [+/-0.7]
Titanic	23.5 [+/-4.7]	23.3 [+/-1.3]	22.6 [+/-1.2]	22.4 [+/-1.0]
German	27.2 [+/-2.7]	24.7 [2.4]	27.5 [2.5]	23.6 [2.1]

Table II
COMPARISON OF RESULTS

REFERENCES

[1] C. J. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, pp. 121–167, 1998.

[2] E.-H. S. Han, G. Karypis, and V. Kumar, "Text categorization using weight adjusted k-nearest neighbor classification," 1999.

[3] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification (2nd Edition)*, 2000.

[4] G. Rtsch, G. R. Atsch, K.-R. Mller, T. Onoda, and K. r. M Uller, "Soft margins for adaboost," in *Machine Learning*, 2000, pp. 287–320.

[5] T. Cooke, "Two variations on fisher's linear discriminant for pattern recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 2, pp. 268–273, 2002.

[6] Y.-C. Hu and J.-F. Tsai, "Evaluating classification performances of single-layer perceptron with a choquet fuzzy integral-based neuron," *Expert Syst. Appl.*, vol. 36, no. 2, pp. 1793–1800, 2009.

[7] A. L. Berger, S. A. D. Pietra, and V. J. D. Pietra, "A maximum entropy approach to natural language processing," *Computational Linguistics*, vol. 22, pp. 39–71, 1996.

[8] S. C. Zhu, Y. N. Wu, and D. Mumford, "Minimax entropy principle and its application to texture modeling," *Neural Comput.*, vol. 9, no. 9, pp. 1627–1660, 1997.

[9] H. Jeffreys, *Theory of Probability (2nd Edition)*. Oxford Clarendon Press, 1948.

[10] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*. New York: Springer-Verlag, 1996.

[11] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K. Mullers, "Fisher discriminant analysis with kernels," in *Neural Networks for Signal Processing IX, 1999. Proceedings of the 1999 IEEE Signal Processing Society Workshop*, Aug 1999, pp. 41–48.