Minimum Description Length Principle for Maximum Entropy Model Selection

Gaurav Pandey and Ambedkar Dukkipati
Department of Computer Science and Automation
Indian Institute of Science, Bangalore 560012
Email: {gaurav.pandey, ad}@csa.iisc.ernet.in

***Abstract*—In maximum entropy method, one chooses a distribution from a set of distributions that maximizes the Shannon entropy for making inference from incomplete information. There are various ways to specify this set of distributions, the important special case being when this set is described by mean-value constraints of some feature functions. In this case, maximum entropy method fixes an exponential distribution depending on the feature functions that have to be chosen a priori. In this paper, we treat the problem of selecting a maximum entropy model given various feature subsets and their moments, as a model selection problem, and present a minimum description length (MDL) formulation to solve this problem. For this, we derive normalized maximum likelihood (NML) code-length for these models. Furthermore, we show that the minimax entropy method is a special case of maximum entropy model selection, where one assumes that complexity of all the models are equal. We extend our approach to discriminative maximum entropy models. We apply our approach to gene selection problem to select the number of moments for each gene for fixing the model.**

I. Introduction

Model selection is central to statistics, and the most popular statistical techniques for model selection include Akaike information criterion (AIC) [1], Bayesian information criterion (BIC) [2], Bayes factors, Bayesian model averaging, cross-validation and minimum description length (MDL) [3]. The basic idea behind MDL principle is to equate compression with finding regularities in the data. Since learning often involves finding regularities in the data, hence learning can be equated with compression as well. Hence, in MDL, we try to find the model that yields the maximum compression for the given observations.

The first MDL code introduced was the two-part code [3]. It was shown that the the two-part code generalizes maximum entropy principle [4]. However, in the past two decades, the normalized maximum likelihood (NML) code, which is a version of MDL, has gained popularity among statisticians. This is particularly because of its minimax properties, as stated in [5], which the earlier versions of MDL did not possess. Efficient methods for computing NML code-lengths for mixture models have been proposed in [6], [7]. In both these papers, the aim of model selection is to decide the optimum number of clusters in a clustering problem.

In a maximum entropy approach to density estimation one has to decide, *a priori*, on the amount of 'information' (for example, number of moments) that should be used from the data to fix the model. The minimax entropy principle [8] states that for given sets of features for the data, one should choose the set that minimizes the maximum entropy. However, it can easily be shown that if there are two feature subsets Φ_1 and Φ_2 and $\Phi_1 \subset \Phi_2$, the minimax entropy principle will always prefer Φ_2 over Φ_1. Hence, though the minimax entropy principle is a good technique for choosing among sets of features with same cardinality, it cannot decide when the sets of features have varying cardinality.

In this paper, we study the NML code-length of maximum entropy models. Towards this end, we formulate the problem of selecting a maximum entropy model given various feature subsets and their moments, as a model selection problem. We derive NML code-length for maximum entropy models and show that our approach is a generalization of minimax entropy principle. We also compute the NML code-length for discriminative maximum entropy models. We apply our approach to gene selection problem for Leukemia data set and compare it with minimax entropy method.

The rest of the paper is organized as follows. In Section II, we briefly discuss about maximum entropy models and NML. In Section III, we derive the NML code-length for maximum entropy models. This is followed by derivation of NML code-length for maximum entropy discriminative models in Section IV. Finally, in Section V, we apply the results obtained in the previous sections for gene selection for Leukemia data set.

II. Preliminaries and Background

Let $\mathcal{X}$ denote the sample space, and $\Delta(\mathcal{X})$ denote the set of all probability distribution on $\mathcal{X}$. $\mathcal{X}^n$ denotes the sample space of all data samples of size n. $\mathbf{x}^n$ denotes an element in $\mathcal{X}^n$, where $\mathbf{x} = (x_1, ..., x_d)$ is a vector in $\mathcal{X}$. MDL principle states that when one has to choose from a set of competing models to explain the data, one should choose the model that leads to best compression of the data. In order to compress the data, we need a code for the data. The two-part MDL code worked by first constructing a code for the parameters, followed by a code for the data given the parameters. A formal introduction to MDL is given in [9]. However, for the sake of completeness, we include some of the material discussed in [9] regarding the NML code, in this paper.

To formalize the definition of normalized maximum likelihood, we need the following notion of 'regret' [9].

978-1-4799-0446-4/13/$31.00 ©2013 IEEE

Definition 1. *Let $\mathcal{M}$ be a model on $\mathcal{X}^n$ and let $\bar{p}$ be a probability distribution on $\mathcal{X}^n$. The regret of $\bar{p}$ with respect to $\mathcal{M}$ for data sample $\mathbf{x}^n$ is defined as*

$$\mathcal{R}(\bar{p}) = -\log \bar{p}(\mathbf{x}^n) - \inf_{p\in\mathcal{M}} [-\log p(\mathbf{x}^n)] \ .$$

The regret is nothing but the extra number of bits needed in encoding $\mathbf{x}^n$ using $\bar{p}$ instead of the optimal distribution for $\mathbf{x}^n$ in $\mathcal{M}$. The worst case regret denoted by $\mathcal{R}_{max}$ is defined as the maximum regret over all sequences in $\mathcal{X}^n$

$$\mathcal{R}_{max}(\bar{p}) = \sup_{\mathbf{x}^n\in\mathcal{X}^n} \left[-\log(\bar{p}(\mathbf{x}^n)) - \inf_{p\in\mathcal{M}} [-\log p(\mathbf{x}^n)] \right] \ .$$

Our aim is to find the distribution $\bar{p}$ that minimizes the maximum regret. To this end, we define the complexity of a model $\mathrm{COMP}(\mathcal{M})$ as

$$\mathrm{COMP}(\mathcal{M}) = \log \int_{\mathbf{x}^n\in\mathcal{X}^n} p(\mathbf{x}^n; \hat{\theta}(\mathbf{x}^n))\, \mathrm{d}\mathbf{x}^n \ , \qquad (1)$$

where $\hat{\theta}(x^n)$ denote the maximum likelihood estimate (MLE) of the parameter θ for the model $\mathcal{M}$ for the data sample x^n. In the above equation and subsequent sections, the integral is defined subject to existence. Also, we assume that MLE is well defined for the model $\mathcal{M}$. Furthermore, the error of a model is defined as

$$\mathrm{ERR}(\mathcal{M}, \mathbf{x}^n) = \inf_{p\in\mathcal{M}} [-\log(p(\mathbf{x}^n)] \ . \qquad (2)$$

The following result is due to [5].

Theorem II.1. *If the complexity of a model is finite, then the minimax regret is uniquely achieved by the normalized maximum likelihood distribution given by*

$$\bar{p}_{nml}(\mathbf{x}^n) = \frac{p(\mathbf{x}^n; \hat{\theta}(\mathbf{x}^n))}{\int_{\mathbf{y}^n\in\mathcal{X}^n} p(\mathbf{y}^n; \hat{\theta}(\mathbf{y}^n))\, \mathrm{d}\mathbf{y}^n} \ .$$

The corresponding code-length $-\log(p_{nml}(\mathbf{x}^N))$ also known as the stochastic complexity of the data sample $\mathbf{x}^n$ is given by

$$\mathrm{NML}(\mathcal{M}, \mathbf{x}^n) = \mathrm{ERR}(\mathcal{M}, \mathbf{x}^n) + \mathrm{COMP}(\mathcal{M}) \ .$$

III. Maximum Entropy Model Selection using MDL

A. Problem Definition

The problem that we intend to address in this paper is as follows.

Let X be a random variable taking values in $\mathcal{X}$, whose probability distribution we wish to estimate. Let $\Phi = \{\phi_1, ..., \phi_m\}$ be a set of functions of random variable whose empirical values have been equated to their expected values in the resultant optimization problem. The resultant linear family $\mathcal{L}_{\Phi,\mathbf{x}^n}$ is given by the set of all probability distributions that satisfy the constraints

$$\int_{\mathbf{x}\in\mathcal{X}} p(\mathbf{x})\phi_k(\mathbf{x})\, \mathrm{d}\mathbf{x} = \bar{\phi}_k(\mathbf{x}^n), \quad 1 \le k \le m \ , \qquad (3)$$

where $\bar{\phi}(\mathbf{x}^n)$ is the empirical estimate of ϕ for the data $\mathbf{x}^n$. The maximum entropy model $\mathcal{M}_{\Phi_l}$ is defined as the set of probability distributions $p \in \Delta(\mathcal{X})$ that have the following form:

$$p(\mathbf{x}) = \exp\left(-\lambda_{0,l} - \sum_{k=1}^{m_l} \lambda_{k,l}\phi_{k,l}(\mathbf{x}) \right) \ , \qquad (4)$$

where $\Lambda = (\lambda_{0,l}, ..., \lambda_{m_l,l}) \in \mathbb{R}^{m_l+1}$. Here, $\lambda_{0,l}$ is the normalizing constant.

Given a set of maximum entropy models characterized by their function set $\Phi_l, 1 \le l \le r$, we use NML code to choose the model that best describes the data.

As a special case, one can choose $\phi(\mathbf{x}) = x_i^j,\ 1 \le i \le d,\ 1 \le j \le m_i,$ for all $\phi \in \Phi$. The aim is to find the best tuple $(m_1, .., m_d) \in \mathbb{N}^d$ for the data. One can further simplify the problem and assume $m_1 = m_2 = ... = m_d = m$. The aim is then to find the best m.

B. NML Code-length

The NML code-length of data $\mathbf{x}^n$ for a given model $\mathcal{M}$ is composed of two parts: (i) the error code-length and (ii) the complexity of the model.

Proposition III.1. *Error code-length of data sequence $\mathbf{x}^n$ for the maximum entropy model $\mathcal{M}_\Phi$ is n times the maximum entropy of the corresponding linear family $\mathcal{L}_{\Phi,\mathbf{x}^n}$*

$$\mathrm{ERR}(\mathcal{M}_\Phi, \mathbf{x}^n) = nH(p^*_{\mathbf{x}^n}), \qquad (5)$$

where $p^*_{\mathbf{x}^n}$ is the maximum entropy distribution of $\mathcal{L}_{\Phi,\mathbf{x}^n}$ given by (3)*.*

Proof: First, we compute the error code-length of the data $\mathbf{x}^n$ for the model $\mathcal{M}_\Phi$. By definition,

$$\mathrm{ERR}(\mathcal{M}_\Phi, \mathbf{x}^n) = \inf_{p\in\mathcal{M}_\Phi} \{-\log(p(\mathbf{x}^n))\} \ .$$

Using definition of $\mathcal{M}_\Phi$ from equation (4), we get

$$\begin{aligned} &\mathrm{ERR}(\mathcal{M}_\Phi, \mathbf{x}^n) \\ &\quad = \inf_{\Lambda} \left[-\sum_{i=1}^{n} \left(-\lambda_0 - \sum_{k=1}^{m} \lambda_k \phi_k(\mathbf{x}^{(i)}) \right) \right] \\ &\quad = n \left[\inf_{\Lambda} \left(\lambda_0 + \sum_{k=1}^{m} \lambda_k \bar{\phi}_k(\mathbf{x}^n) \right) \right] \ , \end{aligned} \qquad (6)$$

where $\bar{\phi}(\mathbf{x}^n)$ is the sample estimate of ϕ for the data $\mathbf{x}^n$ and $\Lambda = (\lambda_0, ..., \lambda_m) \in \mathbb{R}^{m+1}$.

Using Lagrange multipliers, it is easy to see that the maximum entropy distribution for the linear family $\mathcal{L}_{\Phi,\mathbf{x}^n}$ has the form

$$p^*(\mathbf{x}) = \exp\left(-\lambda_0^* - \sum_{k=1}^{m} \lambda_k^* \phi_k(\mathbf{x}) \right) \text{ for all } \mathbf{x} \in \mathcal{X} \ ,$$

for some $\Lambda^* = (\lambda_0^*, ..., \lambda_m^*)$. Since maximum entropy distribution always exists for a linear family, the parameters Λ^* can

be obtained by maximizing the log likelihood function.

$$\Lambda^* = \underset{\Lambda}{\operatorname{argmax}} \left[\sum_{i=1}^{n} \left(-\lambda_0 - \sum_{k=1}^{m} \lambda_k \phi_k(\mathbf{x}^{(i)}) \right) \right] = \underset{\Lambda}{\operatorname{argmin}} \left[\lambda_0 + \sum_{k=1}^{m} \lambda_k \bar{\phi}_k(\mathbf{x}^n) \right] \ . \quad (7)$$

where we remove the negative sign to change argmax to argmin. The notation $\bar{\phi}_k(\mathbf{x}^n)$ is used to denote the empirical mean of $\bar{\phi_k}$ for the data $\mathbf{x}^n$.

The corresponding entropy is given by

$$H(p^*_{\mathbf{x}^n}) = -\sum_{\mathbf{x} \in \mathcal{X}} p^*_{\mathbf{x}^n}(\mathbf{x}^{(i)}) \log(p^*_{\mathbf{x}^n}(\mathbf{x}^{(i)})) = \lambda^*_0 + \sum_{k=1}^{m} \lambda^*_k \bar{\phi}_k(\mathbf{x}^{(i)}) \ , \quad (8)$$

where the last equality follows from the definition of $\mathcal{L}_{\Phi,\mathbf{x}^n}$ in equation (3). By combining equations (7) and (8) and using the fact that maximum entropy distribution always exists for a linear family, we get

$$H(p^*_{\mathbf{x}^n}) = \min_{\Lambda} \left[\lambda_0 + \sum_{k=1}^{m} \lambda_k \bar{\phi}_k(\mathbf{x}^n) \right] \ . \quad (9)$$

By combining equations (9) and (6), we get

$$\mathrm{ERR}(\mathcal{M}_\Phi, \mathbf{x}^n) = nH(p^*_{\mathbf{x}^n}) \ .$$

■

Corollary III.2. *Complexity of the maximum entropy model $\mathcal{M}_\Phi$ is given by*

$$\mathrm{COMP}(\mathcal{M}_\Phi) = \log \int_{\mathbf{y}^n \in \mathcal{X}^n} \exp(-nH(p^*_{\mathbf{y}^n}))\, \mathrm{d}\mathbf{y}^n \ , \quad (10)$$

where $p^*_{\mathbf{y}^n}$ is the maximum entropy distribution of $\mathcal{L}_{\Phi,\mathbf{y}^n}$.

Proof: By definition, we have

$$\mathrm{COMP}(\mathcal{M}_\Phi) = \log \int_{\mathbf{y}^n \in \mathcal{X}^n} \max_{p \in \mathcal{M}_\Phi} p(\mathbf{y}^n)\, \mathrm{d}\mathbf{y}^n \ . \quad (11)$$

Now,

$$\max_{p \in \mathcal{M}_\Phi} p(\mathbf{y}^n) = \exp(\max_{p \in \mathcal{M}_\Phi} \log p(\mathbf{y}^n)) = \exp(-nH(p^*(\mathbf{x}^n))) \ , \quad (12)$$

where we have used the definition of error in (2) and its relationship with entropy in (5) to get the result.

Replacing equation (12) in (11), we get the desired result.

■

Hence, the NML code-length (also known as stochastic complexity) of $\mathbf{x}^n$ for the model M_Φ is given by

$$\mathrm{NML}(\mathcal{M}_\Phi, \mathbf{x}^n) = nH(p^*_{\mathbf{x}^n}) + \log \int_{\mathbf{y}^n \in \mathcal{X}^n} \exp(-nH(p^*_{\mathbf{y}^n}))\, d\mathbf{y}^n \ .$$

C. Generalization Of Minimax Entropy Principle

In this section, we show that the presented NML formulation for maximum entropy is a generalization of the minimax entropy principle [8], where this principle has been used for feature selection in texture modeling.

Let $\Phi_1, ..., \Phi_l$ be sets of functions from $\mathcal{X}$ to the set of real numbers. Corresponding to each set Φ_p, there exists a maximum entropy model $\mathcal{M}_{\Phi_p}$ and vice-versa. For example, if $\Phi_p = \{\phi_{p1}, ..., \phi_{pm_p}\}$, the corresponding maximum entropy model is

$$\mathcal{M}_{\Phi_p} = \left\{ p \in \Delta(\mathcal{X}) : p(\mathbf{x}) = \exp\left(-\lambda_0 - \sum_{k=1}^{m_p} \lambda_k \phi_{pk}(\mathbf{x}) \right) \forall \mathbf{x} \in \mathcal{X} \right\} \ . \quad (13)$$

The MDL principle states that given a set of models for the data, one should choose the model that minimizes the code-length of the data. Here, the code-length that we are interested in is the NML code-length (also known as stochastic complexity). Since, there exists a one-one relationship between the maximum entropy models and the function sets Φ_p, the model selection problem can be reframed as

$$\hat{\Phi} = \underset{\Phi_k, 1 \le k \le p}{\operatorname{argmin}} \left[nH(p^*_{\mathbf{x}^n,k}) + \log \int_{\mathbf{y}^n \in \mathcal{X}^n} \exp(-nH(p^*_{\mathbf{y}^n,k}))\, \mathrm{d}\mathbf{y}^n \right] \ . \quad (14)$$

If we assume that all our models have the same complexity, then the second term in R.H.S can be ignored. Since n, the size of data is a constant, the model selection problem becomes

$$\hat{\Phi} = \underset{\Phi_k, 1 \le k \le p}{\operatorname{argmin}} H(p^*_{\mathbf{x}^n,k}) = \underset{\Phi_k, 1 \le k \le p}{\operatorname{argmin}} \max_{p \in \mathcal{L}_{\Phi_k,\mathbf{x}^n}} H(p) \ .$$

This is the classical minimax entropy principle given in [8]. Hence, the minimax entropy principle is a special case of the MDL principle where the complexity of all the models are assumed to be the same and the models assumed are the maximum entropy models.

IV. Discriminative Models

Discriminative methods for classification, model the conditional probability distribution $p(c|\mathbf{x})$, where c is the class label for data x. Maximum entropy based discriminative classification tries to find the probability distribution with the maximum conditional entropy $H(C|X)$ subject to some constraints [10], where C is the class variable. Initially, information is extracted from the data in the form of empirical means of functions. These empirical values are then equated to their expected values, thereby forming a set of constraints. The classification model is constructed by finding the maximum entropy distribution subject to these sets of constraints. We use MDL to decide the amount of information to extract from the data in the form of functions of features. A straightforward application of this technique is feature selection.

The maximum entropy discriminative model $\mathcal{M}_\Phi$, where $\Phi = \{\phi_1, ..., \phi_m\}$ is the set of all probability distributions of the form

$$p(c|\mathbf{x}) = \frac{\exp(-\sum_{k=1}^m \lambda_k \phi_k(\mathbf{x}, c))}{\sum_{c\in\mathcal{C}} \exp(-\sum_{k=1}^m \lambda_k \phi_k(\mathbf{x}, c))} \ ,$$

Let us denote the denominator in above equation as $Z(\mathbf{x})$. Let $\mathcal{L}_{\Phi,\mathbf{x}^n,c^n}$ be the linear family given by

$$\mathcal{L}_{\Phi,\mathbf{x}^n,c^n} = \Bigg\{ p \in \Delta(\mathcal{X}) : \int_{\mathbf{x}\in\mathcal{X}} \sum_{c\in\mathcal{C}} p(c|\mathbf{x}) p(\mathbf{x}) \phi_k(\mathbf{x}, c)\, \mathrm{d}\mathbf{x} = \frac{1}{n} \sum_{i=1}^n \phi_k(\mathbf{x}^{(i)}, c^{(i)}) \ , 1 \le k \le m \Bigg\} \ .$$

Since, we are not interested in modelling $p(\mathbf{x})$, we use the empirical distribution $\tilde{p}(\mathbf{x})$to approximate $p(\mathbf{x})$ as done in [11]. The empirical distribution $\tilde{p}(\mathbf{x})$ is given by

$$\tilde{p}(\mathbf{x}) = \begin{cases} \frac{1}{n} & \mathbf{x} \in \mathbf{x}^n \\ 0 & \text{otherwise} \ . \end{cases} \tag{15}$$

Hence, the constraints become

$$\sum_{i=1}^n \sum_{c\in\mathcal{C}} p(c|\mathbf{x}^{(i)}) \phi_k(\mathbf{x}^{(i)}, c) = \sum_{i=1}^n \phi_k(\mathbf{x}^{(i)}, c^{(i)}) \ . \tag{16}$$

As discussed in [12], the sender-receiver model assumed here is as follows. Both sender and receiver have the data $\mathbf{x}^n$. The sender is interested in sending the class labels c^n. If he sends the class labels without compression, he needs to send $n \log(|C|)$ bits. If, however, he uses the data to compute a probability distribution over the class labels, and then compress c^n using that distribution, he may get a shorter code-length for c^n. His goal is to minimize this code-length, such that the receiver can recover the class labels from this code.

We state without proof the following results.

Proposition IV.1. *The Error code-length of $c^n|\mathbf{x}^n$ for the conditional model $\mathcal{M}_\Phi$ is equal to n times the maximum conditional entropy of the model $\mathcal{L}_{\Phi,\mathbf{x}^n,c^n}$.*

Corollary IV.2. *The complexity of the conditional model $\mathcal{M}_\Phi$ is given by*

$$\mathrm{COMP}(\mathcal{M}_\Phi) = \sum_{y^n \in \mathcal{C}^n} \exp(-nH(p^*_{y^n})) \ .$$

V. Application to Gene Selection

We use gene selection as an example to illustrate discriminative model selection for maximum entropy models. The dataset used is Leukemia dataset available publicly at http://www.genome.wi.mit.edu. The data set consists of two classes: acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL). There are 38 training samples and 34 independent test samples in the data. The data consists of 7129 genes. The genes are preprocessed as recommended in [13].

Assuming the sender-receiver model discussed above, the sender needs 38 bits or 26.34 nats in order to send the class labels of training data to receiver. If the NML code is used, the sender needs 24.99 nats. Since the sender and receiver both contain the microarray data, the sender can use the microarray data to compress the class labels much more than can be obtained without the microarray data. Specifically, we are interested in finding the genes which gives the best compression, or the minimum NML codelength.

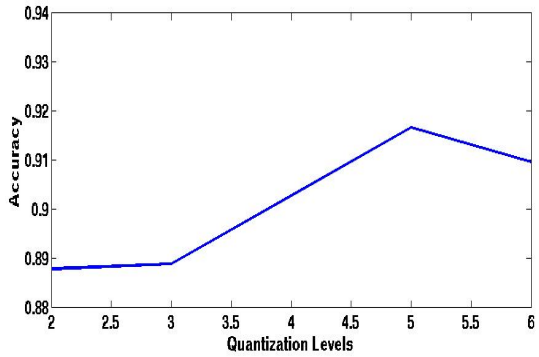

Fig. 1. Change in accuracy with quantization level for the top 25 genes

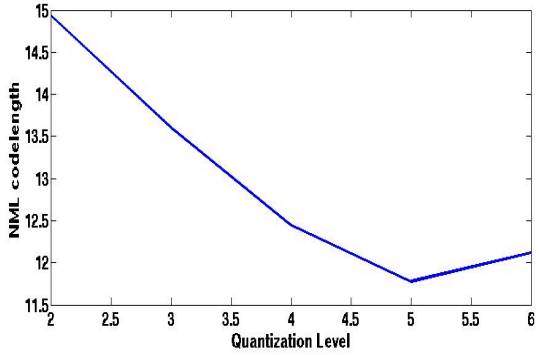

Fig. 2. Change in average NML code-length with quantization level for the top 25 genes

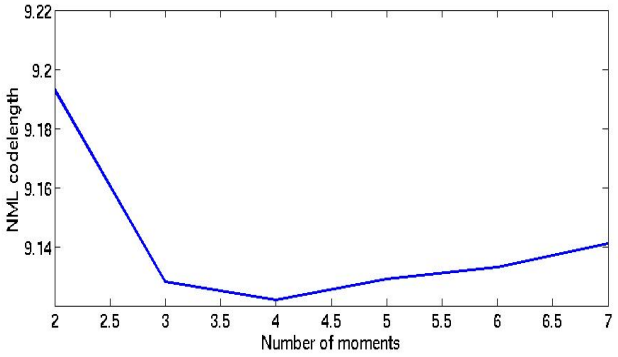

Fig. 3. Variation of NML code-length of class labels for gene M16038 with the number of moment constraints m

For the purpose of our algorithm, we quantize the genes to various levels. We claim that quantizing a gene reduces the

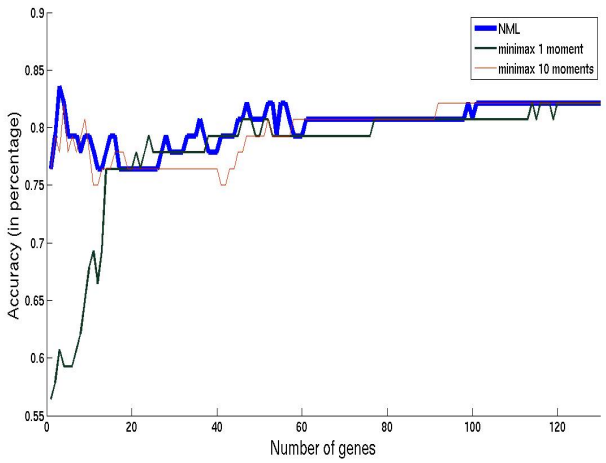

Fig. 4. Comparison of accuracy of various maximum entropy classifiers based on NML and minimax entropy.

risk of overfitting of the model to data. To support our claim, we have also plotted the change in accuracy with quantization level in Figure 1 for the top 25 genes. As can be seen from the graph, increasing quantizaton level from 5 to 6 results in a decrease in accuracy. We have also plotted a graph for change in average NML code-length with quantization level for the top 25 genes in Figure 2. An interesting observation is that the minima of NML code-length coincides exactly with the maxima of accuracy. A similar trend was observe when the number of genes were changed.

Hence, we quantize each gene to 5 levels. Other than the advantages of quantization mentioned above, quantization is also necessary for the current problem as the problem of calculating complexity can become intractable even for moderate n. The constraints that we use are moment constraints, that is $\phi_k(\mathbf{x}) = \mathbf{x}^k, 1 \leq k \leq m$. We vary the value of m from 1 to 7 to get a sequence of maximum entropy models. The NML code-length of the class labels is calculated for each such model. The model that results in the minimum NML code-length is selected for each gene.

It was observed that for most genes, the NML code-length decreased sharply when m was increased from 1 to 2. The change in values of NML code-length was less noticeable for $m \geq 2$. The variation of NML code-length of class labels for a typical gene are shown in figure 3. In order to make the changes in NML code-length more visible, we skip the NML code-length for m=1.

Our approach for ranking genes is as follows. For each gene, we select the value of m that gives the minimum NML code-length. We then sort the genes in increasing order of their minimum NML code-lengths. The minimum code-length achieved is 8.35 nats, which is much smaller than 24.99 nats achieved without using the microarray data. Since compression is equated with finding regularity according to Minimum Description Length principle, hence, it can be stated that the topmost gene is able to discover a lot of regularity in the data.

Finally, we use MDL to build a classifier. The amount of information to use for each gene is decided, by using MDL to fix the number of moments. MDL is used to rank the features. Then, we use class conditional independence among features to build a maximum entropy classifier. The number of genes used for the classifier are varied from 1 to 130. The resultant graph is compared with other maximum entropy classifiers in Figure 4, where the amount of information used per gene is the same for all genes.

VI. Conclusion

Finding appropriate feature functions and the number of moments is important to any maximum entropy method. In this paper, we pose this problem as a maximum entropy model selection problem and develop an MDL based method to solve this problem. We showed that this approach generalizes minimax entropy principle of [8]. We derived NML code-length in this respect, and extended it to discriminative maximum entropy model selection. We tested our proposed method for gene selection problem to decide on the quantization level and number of moments for each gene. Finally, we selected the genes based on the code-length of the class labels and compared the simulation results with minimax entropy method. The bottleneck for using MDL for model selection in discriminative classification is the computation of complexity. More efficient approximations to calculate the complexity need to be developed to employ this approach for problems involving larger data sets.

References

[1] H. Akaike, "A new look at the statistical model identification," *Automatic Control, IEEE Transactions on*, vol. 19, no. 6, pp. 716–723, 1974.
[2] G. Schwarz, "Estimating the dimension of a model," *The annals of statistics*, vol. 6, no. 2, pp. 461–464, 1978.
[3] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, no. 5, pp. 465–471, 1978.
[4] M. Feder, "Maximum entropy as a special case of the minimum description length criterion (corresp.)," *Information Theory, IEEE Transactions on*, vol. 32, no. 6, pp. 847–849, 1986.
[5] Y.M. Shtar'kov, "Universal sequential coding of single messages," *Problemy Peredachi Informatsii*, vol. 23, no. 3, pp. 3–17, 1987.
[6] P. Kontkanen and P. Myllymäki, "Computing the regret table for multinomial data," 2005.
[7] S. Hirai and K. Yamanishi, "Efficient computation of normalized maximum likelihood coding for gaussian mixtures with its applications to optimal clustering," in *Information Theory Proceedings (ISIT), 2011 IEEE International Symposium on*. IEEE, 2011, pp. 1031–1035.
[8] S.C. Zhu, Y.N. Wu, and D. Mumford, "Minimax entropy principle and its application to texture modeling," *Neural Computation*, vol. 9, no. 8, pp. 1627–1660, 1997.
[9] P. Grunwald, "A tutorial introduction to the minimum description length principle," *Arxiv preprint math/0406077*, 2004.
[10] A.L. Berger, V.J.D. Pietra, and S.A.D. Pietra, "A maximum entropy approach to natural language processing," *Computational linguistics*, vol. 22, no. 1, pp. 39–71, 1996.
[11] K. Nigam, J. Lafferty, and A. McCallum, "Using maximum entropy for text classification," in *IJCAI-99 workshop on machine learning for information filtering*, 1999, vol. 1, pp. 61–67.
[12] I. Tabus, J. Rissanen, and J. Astola, "Classification and feature gene selection using the normalized maximum likelihood model for discrete regression," *Signal Processing*, vol. 83, no. 4, pp. 713–727, 2003.
[13] S. Dudoit, J. Fridlyand, and T.P. Speed, "Comparison of discrimination methods for the classification of tumors using gene expression data," *Journal of the American statistical association*, vol. 97, no. 457, pp. 77–87, 2002.