Analyzing the Incomplete Data Based on the Improved Maximum Entropy Model

Jian Zhao, Xiao-Long Wang, Yi Guan, Lei Lin

School of Computer Science and Technology,
Harbin Institute of Technology, 150001 Harbin, P.R.

{zhaojian, wangxl, guangyi, linl}@insun.hit.edu.cn

Abstract

When MEM (Maximum Entropy Model) trained by GIS (Generalized Iterative Scaling) algorithm was used to analyze the incomplete data, in order to satisfy the constraint of GIS, a global unique compensating feature was introduced to offset the effect of missing attributes of some samples on classification result. However, this kind of compensating strategy neglected a basic fact that different features had different effect on classification result. In this paper, an improved compensating strategy was proposed to overcome the shortage of traditional method: took effects of both different feature types and label types into account. Experiment results on Mushroom data set coming from UCI data repository showed that the new method was feasible and effective. The average error rate was reduced by about 68.3% and 33.5% respectively on two kinds of dataset.

Keyword: maximum entropy model, incomplete data, features compensating, pattern classification.

I. Introduction

Most information was presented as image, voice, text, or data warehouse, which is always unstructured. How to extract useful knowledge from these data is the researching hotspot of pattern recognition and machine learning domain. The issue of analysis of incomplete data was firstly proposed by Hartley in 1971 [1], Granger [2] has summarized four types of incomplete data:

1) Incompleteness caused by data sparseness;
2) Incompleteness because of some attribute of samples were missing;
3) Incompleteness induced by missing class labels of some samples were;
4) Incompleteness induced by missing classes. Some classes that are not present in the training set may be encountered during testing.

Maximum entropy model is a common method which is always used to do classification or pattern recognition. MEM was firstly proposed by Jaynes in 1950’s [3]. It can be adopted to process the incomplete data, especially suitable for first two kinds of incompleteness. Recently MEM was utilized to do the research of text mining, such as text classification [4], part of speech tagging [5], named entity recognition [6], and so on. In theses MEM’s applications, incompleteness of data sparseness is more common. Some research (mainly on the data smoothing technology) has been conducted to make MEM overcome this kind of incompleteness [7]. Additionally MEM can also be used to perform the data mining task. Arne designed a system which using MEM to combine several naïve Bayesian models, and this system achieved good result in Data-Mining-Cup 2004[8]. Raychaudhuri tested the performance of MEM, naïve Bayesian model, and K-nearest neighbor method in text classification, and drew a conclusion that the MEM outperform the other two methods, then he used MEM for genes function tagging [9]. Dong Qiwen utilized the MEM integrated with word lattice techniques for protein secondary structure prediction [10], the precision of the MEM exceeded the classic neural networks. In such applications, the second type of incompleteness is more marked. However, there are hardly any discussions about how to process data missing some attribute.

The principle of maximum entropy asserts that the only probability distribution that can justifiably be constructed from incomplete information, such as finite training data, is that which has maximum entropy subject to a set of constraints representing the information available. GIS used to be the main parameter estimation method for MEM. As for the classification of incomplete data running short some attributes, in order to satisfy the constraint of GIS, a global unique compensating feature was introduced to offset the effect of missing attributes of some samples on classification result. However, this kind of compensating strategy neglected a basic fact that different features had different effect on classification result. In this paper, a new compensating strategy was proposed, which took effects of both different feature types and label types into account, to overcome the shortage of traditional compensating method, especially for classifying the second type of incomplete data.

The remnant content of the paper is organized as following: classification of data with missing some attributes, together with basic maximum entropy principle, is introduced in part two. In part 3, we presented the improved feature compensating strategy. Experiment results and analysis were given in part four. Conclusion of the paper was drawn in last part.

II. Simple Introduction of classification of Incomplete Data and Maximum Entropy Principle

A. *The Formal Description of Incomplete Data Classification*

Without losing generality, we just referred to data with missing attributes as incomplete data in the following paper. The formal description of incomplete data classification was given as following.

Supposed that a sample marked as s_i had k types of attribute, and then a complete sample can be denoted as $s_i = \{x_{i1}, \cdots, x_{ik})$. If there are two attributes $i, j (0 \le i < j \le k)$ missing values, then the corresponding sample can be denoted as:

$$s_i = \{x_{i1}, \cdots, x_{i-1}, y_{i1}, x_{i+1}, \cdots, x_{j-1}, y_{i2}, x_{j+1}, \cdots, x_{ik}\} \tag{1}$$

Where y_{i1} and y_{i2} are lost attributes. If all the lost attributes were marked as y_i and observed attributes were marked as x_i, the sample s_i can be rewritten as:

$$s_i = \{x_i, y_i\} \tag{2}$$

Then the whole sample space is $S = \{s_1, \cdots, s_n\} = \{X, Y\}$. In conditional probabilistic model, the task of classifying the incomplete data is to find a class label which can maximize the conditional probability

$$c = \arg\max_{c_j \in C} p(c_j \mid s_i) = \arg\max_{c_j \in C} p(c_j \mid (x_i, y_i)) \tag{3}$$

So the task of classification is to estimate the probability density.

B. *Maximum Entropy Principle*

MEM is a common classification method, a kind of exponential model. It originally rooted in the problem of conditional function extremum. The target function, i.e. conditional information entropy is defined as following [11]:

$$H(p) = -\sum \tilde{p}(s)p(c \mid s)\log p(c \mid s) \tag{4}$$

Where c is class label, s is sample to be classified。The desired variable is $p(c \mid s)$ in equation (3). The conditions are:

$$P = \{p \mid E_p f_i = E_{\tilde{p}} f_i, 1 \le i \le k\} \tag{5}$$

$$\sum_j p(c_j \mid s_i) = 1 \tag{6}$$

Where $E_{\tilde{p}} f_i$ denotes the empirical expectation of f_i, $E_p f_i$ denotes the model expectation of f_i. They are defined as follows:

$$\begin{aligned} E_p f_i &= \sum_{c,h} \tilde{p}(s)\, p(c \mid s)\, f_i(c,s) \\ E_{\tilde{p}} f_i &= \sum_{c,h} \tilde{p}(c,s)\, f_i(c,s) = \frac{1}{N}\sum f_i(c,s) \end{aligned} \tag{7}$$

$$f_i(c,s) = \begin{cases} 1 & if\ : c = c' \ and\ \ h(s) = TRUE \\ 0 & else \end{cases} \tag{8}$$

Here $h(s)$ is predicate function. Through Lagrange transformation on (5), (6) and (7), the expression of $p(c \mid s)$ can be inferred as

$$p(c \mid s) = \frac{1}{Z(s)}\exp\left(\sum_i \lambda_i f_i(c,s)\right) \tag{9}$$

$$Z(s) = \sum_c \exp\left(\sum_i \lambda_i f_i(c,s)\right) \tag{10}$$

GIS is a commonest method for solving the λ_i in above equations. GIS has a constraint that the number of attributes in every sample is equal [11]:

$$M = \max_{s \in S} \sum_{j=1}^{k} f_j(c, h(s)) \tag{11}$$

So as for the task of classification of incomplete data, in order to fulfill the GIS's constraint, a global unique compensating feature was introduced to offset the effect of missing attributes on classification result:

$$\forall s \in S \quad f_{comp}(c, h(s)) = M - \sum_{j=1}^{k} f_j(c, h(s)) \tag{12}$$

As to s_i missing two attributes mentioned in 2.1 section, the traditional MEM treat the y_{i1} and y_{i2} with no difference, i.e. different lost attributes having the same effect on different classes. So the s_i can be rewritten like this

$$s_i = \{x_{i1}, \cdots, x_{i,i-1}, x_{i,i+1}, \cdots, x_{i,j-1}, x_{i,j+1} \cdots, x_k, 2y_i\} \tag{13}$$

The feature of MEM is correlative with target class label, so the features space of sample s_i when it is classified as c_l can be denoted as follows:

$$\begin{aligned} F(s_i, c_l) = \{ & f_{i,1}(x_{i1}, c_l), \cdots, f_{i,i-1}(x_{i,i-1}, c_l), f_{i,i+1}(x_{i,i+1}, c_l), \cdots, \\ & f_{i,j-1}(x_{i,j-1}, c_l), f_{i,j+1}(x_{i,j+1}, c_l), \cdots, f_{i,k}(x_{i,k}, c_l), 2 f_{comp} \} \end{aligned} \tag{14}$$

Then weights of all features were computed by GIS algorithm. However, the traditional feature compensating strategy has following two shortages

- Did not distinguish the difference between different feature types;
- Did not take the effect of compensating features on different target classes into account.

III. Improved Features Compensating Strategy

In order to estimate the weight of different compensating features, the experiential expectation of these features must be computed firstly. As for the expectation of the i th kind of feature for the j th class, it can be computed as follows:

```
Input: samples set $S = \{s_1, \cdots, s_n\}$, classes set $C = \{c_1, \cdots, c_L\}$
Output: parameters of MEM, i.e. $\lambda = \{\lambda_1, \cdots, \lambda_T\}$
1. Form the event space based on samples $EVE = \{e_1, \cdots, e_m\}$   $m \le n$;
2. Add the compensating features array COM_F[1 → k][1 → L] , initialize them to be 0;
3. Compute the expectation value of all features including compensating features;
4. Set up the initial model with $\lambda^0 = \{\lambda_1^0 = 0, \cdots, \lambda_T^0 = 0\}$ , p_old=p_new=0.0;
5. while ( (p_new-p_old)>0 );
        1) p_old=p_new;
        2) for i=1 to i=m
               for j=1 to j=k
                 for n=1 to n=l
                       3) if $e_i$ is defect of jth feature when it was classed as
                              nth category , add COM_F[j][n] to $e_i$ ;
                       4) compute the probability of $e_i$ belonging to nth class
                          using equation (9) with $\lambda^{old}$ ;
               5) count the events which were classified correctly;
               6) compute the precision of this model, i.e. p_new;
               7) update the model expectation of all features;
        8) update the new model using the equation as follows.
           $\lambda_j^{new} = \lambda_j^{old} + {}^1\!/_k (\log E_{\tilde{p}} f_j - \log E_p f_j)$;
        9) $\lambda^{old} = \lambda^{new}$ ;
```

Figure 1: new algorithm for estimating the MEM parameters

$$f_{i,j}^{comp}(c, h(s)) = \begin{cases} 1 & if\ c = c_j\ and\ h_i(s) = false \\ 0 & \end{cases} \tag{15}$$

$$E_{\tilde{p}} f_{i,j}^{comp} = \frac{count\,(sample_{i,j})}{N} \tag{16}$$

Where N is total number of samples, and $count(sample_{i,j})$ is the number of samples which missed the i th feature and belonged to the j th category.

It should be noted that unlike the traditional compensating feature f_{comp} in equation (13) which ranged from 0 to M, the new compensating feature $f_{i,j}^{comp}$ was binary value function just like ordinary features. The new parameters estimating algorithm was given as Fig. 1.

Ⅳ. Experiment and Analysis

A. Experimental Design

The data for experiment is Mushroom data set, coming from UCI machine learning data repository [12]. There are totally 8124 samples with every sample containing 22 kinds of attributes. Among these samples, there are 2480 samples that had no observation value on the eleventh attribute. We selected 2124 from those incomplete samples as testing set, and the other 6000 samples as training set. In order to test the influence of different types of unobserved attribute on classification result, some special training sets were constructed by deleting one or more attributes of samples manually, which were listed in table 1.

Table 1：the data sets for expriment

Data Sets Label	**Volume**	**ID of Missing Attribute**	**The Number of Missing Attributes**
Test	2124	11	2124
Train2	6000	1	356
⋮	⋮	⋮	⋮
Train10	6000	11	356
Train11	6000	1—2，11	500
⋮	⋮	⋮	⋮
Train17	6000	1—8，11	500

In the following content, we referred to the MEM with traditional feature compensating strategy as C1, and the MEM with improved feature compensating strategy as C2.

Table 2: the compensating methods adopted in experiments

ID	**Compensating strategy**	**Number of Comp. features**
C1	Gloablly and uniquely compensating	1
C2	Compensating depended on feature types and target class	22×2

There are four evaluating standard in our experiment: precision (PR), error rate (ER), learning rate (LR) and error declining rate (EDR), which were defined respectively as following:

$$PR = \frac{count(correct)}{total} \qquad ER = 1 - PR \tag{17}$$

$$LR = \frac{precision_{stop}}{count(stop)} \tag{18}$$

$$EDR = \frac{ER_{new} - ER_{old}}{ER_{old}} \tag{19}$$

Where $count(correct)$ — the number of samples classified correctly;

$total$ — the total number of samples;

$precision_{stop}$ — the precision at the time when iterative algorithm stopped;

$count(stop)$ — the number of steps that the GIS required for convergence;

ER_{new} —the error rate of improved new model C2;

ER_{old} — the error rate of original model C1.

Furthermore we used average information entropy of features to evaluate the variety of feature's output:

$$H(f)_{avg} = \frac{H(f)}{L} = \frac{-\sum_{i=1}^{L} p_i \log p_i}{L} \qquad (20)$$

AIE (Average Information Entropy) values of some features were listed in table 3.

Table 3: AIE values of some features

Feature Type	Entropy	Number of Output classes	AIE Value (*10)
1	0.4236	6	0.706
2	0.4637	4	1.159
3	0.7718	10	0.772
4	0.295	2	1.475
5	0.6216	8	0.77
6	**0.0094**	**2**	**0.047**
7	0.204	2	1.02
8	0.1891	2	0.946
11	0.4946	4	1.237

Two kinds of experiments were designed: the first one is for testing the performance of C1 and C2 on dataset with every training sample just losing only one feature. The corresponding training sets were Train2～Train10 in table 1, and the testing set was unique. The second experiment is for testing the performance of C1 and C2 on dataset with every training sample losing more than one features. Seven feature sets ({1-2，11}～{1-8，11}) were chose as testing collect, and the corresponding training sets were Train11～Train17.

B. *Experimental Result and Analysis*

Table 4 showed the performance of C1 and C2 on datasets with missing only one feature. The overall comparison of average precision, average error rate, and average error declining rate of C1 and C2 on Train2～Train9 were listed in table 5.

Tale 4：the performance of C1 and C2 on datasets with missing one feature

Data set	C1		C2	
	PR (%)	LR (%)	PR (%)	LR (%)
Train2	95.5	5.86	98.7	3.18
Train3	95.2	6.8	99.1	3.41
Train4	95.0	3.98	98.9	5.8
Train5	95.3	3.84	99.0	3.8
Train6	94.5	3.69	99.76	12.2
Train7	95.3	4.75	**94.8**	3.65
Train8	95.4	4.99	99.1	3
Train9	95.1	4.83	98.4	3.54
Train10	94.2	5.86	98.1	3.94

Table 5：overall comparison of C1 and C2 on Train2～Train10

Compensating method	Average LR	Average ER	Average EDR
C1	4.93	4.95	——
C2	4.87	1.57	0.683

It can be observed from table 4 that, as for the datasets with missing only one feature, the C2 method outperformed the C1 method, except that on the train7 missing the sixth feature, the classification precision of C2 is lower than C1. The table 5 showed that the improved compensating method C2 reduced the average error rate by 68.3% (from 4.95% to 1.57%), at the same time the average learning rate was almost kept invariable (only 0.06% variance).

In table 6, the learning rate and classification precision of C1 and C2 on datasets with every sample in training set losing more than one attributes were listed. Table 7 showed the overall performance of C1 and C2 on Train11～Train17. In Table 8, the performance comparison between three models, i.e. baseline model (naïve Bayes model), traditional maximum entropy model (C1), and improved maximum entropy model (C2), are listed.

Table 6：performance of C1 and C2 on Train11～Train17

Data set	LR (%)		PR (%)	
	C1	C2	C1	C2
Train11	4.53	4.27	96.4	98.9
Train12	3.98	5.18	97.1	98.7
Train 13	5.84	7.66	98.0	98.7
Train14	3.67	13.93	97.8	99.7
Train15	6.18	13.93	97.4	**94.6**
Train16	2.76	13.93	98.0	**97.5**
Train17	3.54	12.2	98.1	**97.5**

Table 7：the overall performance of C1 and C2 on Train11～Train17

Compensating method	Average LR	Average ER	Average EDR
C1	4.36	3.1	——
C2	10.16	2.06	0.335

Tale 8：the performance of baseline model (naive Bayes model), C1 and C2

Data set	Naïve Bayes	C1	C2
Train2～Train10	91.68	95.05	98.43
Train11～Train17	85.75	96.9	97.94

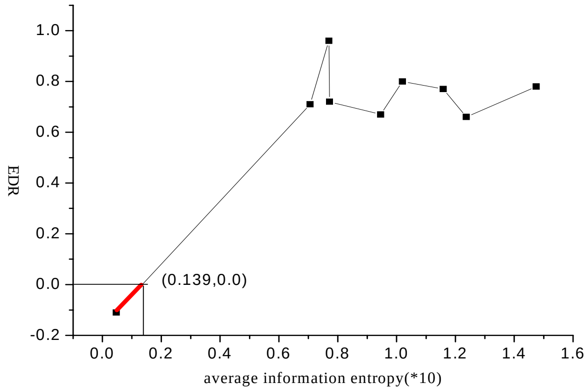

Figure 2：The relation curve of EDR to average information entropy

It can be seen from table 6 that C2 outperformed C1 on most datasets in which samples lost more than one attributes, except on Train14～Train16. These three training sets had one common point that they all lost the sixth feature, which is same as Train7. The datum in table 7 showed that C2 increased the learning rate by more than two times, and the average error declining rate decreased by 33.5%. From Table 8, we can see that both C1 and C2 outperform the baseline model, i.e. naïve Bayes model on two kinds of dataset.

As to the training datasets in which the sixth attribute had no observation value (Train7，Train14～Train16), the precision of improved compensating method C2 is lower than traditional method C1. Through analyzing the composing of every kind of attributes, we found that the performance of improved compensating method proposed in this paper is relative to the AIE value of unobserved attribute. As figure 2 showed, only when the AIE value of unobserved attribute is above a certain value (about 0.014 showed in Fig. 2), the classification error rate of C2 is lower than C1. However, the AIE value of the sixth attribute almost tends to be zero (0.0047 showed in table 3), that is to say this kind of attribute has a too low discriminative ability for classification, and this is opposite to the potential assumption of the new compensating method that different attribute types have different effect on various target classes. So the performance of C2 on datasets excluding the sixth attributes is worse than C1. It is considered reasonable to combine the traditional compensating strategy and improved compensating method to cope with the above case.

Ⅴ. Conclusion

Maximum entropy model can be used to classify the incomplete data. As for the GIS algorithm which is always utilized to estimate the parameters of MEM, there is a constraint that the every sample in datasets must have the same number of attributes. In order to fulfill this constraint, the traditional method used a global and unique feature to substitute the lost attribute. In this paper, an improved compensating method was proposed to overcome the shortage of traditional method that ignored the effect of different attribute types on different class labels. The results of experiment on Mushroom dataset showed that the new method outperformed the traditional one: the new method decreased the average error rate by 68.3% on dataset missing only one kind of attribute, and 33.5% on dataset missing more than one kinds of attribute. Furthermore, it was discovered that the classification result of new compensating method is depended on the average information entropy of unobserved attributes. If the value of AIE was too small, i.e. tend to be zero, the performance of new method will not be as good as that of traditional method.

Acknowledgements

This investigation was supported by the National Natural Science Foundation of China (60435020) and The 863 high Technology Research and Development Programme of China (2002AA117010-09)

References

[1] Hartley. H.& Hocking. R. The analysis of incomplete data. Biometrics, 1971. Vol. 27, 783-808.

[2] Granger, E.; Rubin, M.A.; Grossberg, S.; Lavoie, P. Classification of incomplete data using the fuzzy ARTMAP neural network. 2000. IJCNN 2000, Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. July 2000, Vol. 6, 24-27.

[3] E.T. Jaynes. Information Theory and Statistical Mechanics. Physics Reviews. 1957, vol.106: 620-630.

[4] Kamal Nigam, John Lafferty, et al. Using maximum entropy for text classification. In proceedings of the IJCAI-99 workshop on information filtering, Stockholm, SE, 1999.

[5] Zhao Jian , Xiao-long Wang, Chinese POS Tagging based on Maximum Entropy Model. IEEE Proceedings of 2002 International Conference on Machine Learning and Cybernetics, Beijing, Vol.1. 2002.

[6] Andrew Borthwick, A maximum entropy approach to named entity recognition, PhD dissertation New York University, September,1999.

[7] S. Chen and R. Rosenfeld, A Gaussian Prior for Smoothing Maximum Entropy Models, Computer. Sci. Dept., Carnegie Mellon Univ., Pittsburgh, PA, Tech. Rep. CMU-CS-99-108, Feb. 1999.

[8] Arne Mauser. http://www.prudsys.de/Service/Downloads/bin/dmc2004_report_eng.pdf. Technical report .

[9] Soumya Raychaudhuri,1 Jeffrey T. Chang, *et al.* Associating Genes with Gene Ontology Codes Using a Maximum Entropy Analysis of Biomedical Literature. Genome Research. 2002 Jan. Vol. 12, Issue 1, 203-214

[10] Dong Qinwen, Wang Xiaolong, Lin Lei, *et al.* A Seqlet-based Maximum Entropy Markov Approach for Protein Secondary Structure Prediction. Science in China Ser. C Life Sciences, 2005, 35 (1): 87~96.

[11] Adwait Ratnaparkhi. A Simple Introduction to Maximum Entropy Models for Natural Language Processing. Technical Report. May 1997.

[12] Jeff Schlimmer. UCI Repository of machine learning databases. Irvine, CA: University of California, http://www.ics.uci.edu/~mlearn/MLRepository.html, 1987.

Jian Zhao is a PhD candidate in School of Computer Science and Technology of Harbin Institute of Technology. He received the Bachelor degree from ChongQing University, P.R. China in 1997, and received the Master degree from KunMing University of Science and Technology, P.R. China in 2000. His current research interests include machine learning, information retrieval, natural language processing and information extraction.

Xiaolong Wang received the B.E. degree in computer science from Harbin Institute of Electrical Technology, China, and the M.E. degree in Computer Architecture from Tianjin University, China, in 1982, and 1984, respectively, and the Ph.D. degree in Computer Science and Engineering from Harbin Institute of Technology, China, in 1989. He

GUAN YI is presently an associate professor of the School of Computer Science and Technology at Harbin Institute of Technology. He received the Bachelor degree in Computer Science and Technology from Tianjin University in 1992. He obtained the Master and PhD. degrees in Computer Science and Technology from Harbin Institute of Technology in 1995 and 1999 respectively. His research interests include question answering, statistical language processing, parsing, and text mining.

Lei Lin is presently a lecturer of the School of Computer Science and Technology at Harbin Institute of Technology. He received the Ph.D. degree in Computer Science and Technology form Harbin Institute of Technology in 2004. His research interests include protein structure prediction, Gene expression data analysis, Gene region prediction, and Chinese handwritten characters recognition post-processing and information fusion.