A Semi-Supervised Regression Model for Mixed Numerical and Categorical Variables *

Michael K. Ng† Elaine Y. Chan‡ Meko M.C. So§ Wai-Ki Ching ¶

Abstract

In this paper, we develop a semi-supervised regression algorithm to analyze data sets which contain both categorical and numerical attributes. This algorithm partitions the data sets into several clusters and at the same time fits a multivariate regression model to each cluster. This framework allows one to incorporate both multivariate regression models for numerical variables (supervised learning methods) and k-modes clustering algorithms for categorical variables (unsupervised learning methods). The estimates of regression models and k-modes parameters can be obtained simultaneously by minimizing a function which is the weighted sum of the least squares errors in the multivariate regression models and the dissimilarity measures among the categorical variables. Both synthetic and real data sets are presented to demonstrate the effectiveness of the proposed method.

Key Words: Clustering, regression, data mining, numerical variables, categorical variables

*The research is supported in part by Hong Kong Research Grants Council Grant Nos. 7046/03P, 7035/04P, 7035/05P and HKBU FRGs.

†Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong, email: mng@math.hkbu.edu.hk

‡Department of Mathematics, The University of Hong Kong, Pokfulam Road, Hong Kong

§School of Management, The University of Southampton, Highfield, Southampton, SO17 1BJ, United Kingdom. email: mekoso@soton.ac.uk

¶Department of Mathematics, The University of Hong Kong, Pokfulam Road, Hong Kong. email:wching@hkusua.hku.hk

1 Introduction

Regression analysis is a statistical technique that allows one to assess the relationship between a dependent variable and several independent variables. As a popular application in many disciplines, the use of regression analysis has been widely discussed in literatures [17, 20, 21, 16, 23]. In the classical regression analysis, the dependent variable and the independent variables fall into a continuous numeric domain and all statistical operations are terminologically computed on the set of real numbers. If the given dependent variable is a set of dichotomous data, the accuracy of regression analysis becomes sensitive to the size of the independent variables [18].

Logistic regression has been developed to predict a discrete outcome from a set of variables that may be continuous, discrete, dichotomous, or a mix. In logistic regression, the dependent variable can take the value 1 with a probability of success θ, or the value 0 with probability of failure 1-θ. By using the logistic function,

$$\log\left(\frac{\theta}{1-\theta}\right) = \exp(\alpha + \beta_1 x_1 + \cdots + \beta_i x_i)$$

where x_i are independent variables, α and β_i are parameters to be estimated in the model. The goal of logistic regression is to correctly predict the category of outcome for individual cases using the most parsimonious model. For instance, Mlogit techniques [18, 23] use the multinomial logit distribution to model unordered categorical variables. The dependent variable may be in the format of either character strings or integer values. A general class of regression models for ordinal data is developed in [17], for instance PLUM (Polytomous Universal Model). These models utilize the ordinal nature of the data by describing various modes of stochastic ordering and this eliminates the need for assigning scores or otherwise assuming cardinality instead of ordinality. Two models in particular, the proportional odds and the proportional hazards models are likely to be most useful in practice because of the simplicity of their interpretation. These linear models are shown to be multivariate extensions of generalized linear models. Hubert and Rousseeuw [13] explored the use of regression models where the regressors are both numerical and ordinal data. They introduced a robust regression method which performs a weighted least absolute values fit to both types of data. In [9], Hathaway and Bezdek studied switching regression models and fuzzy clustering for numerical data.

A simple regression model may not be able to handle the need of today's information abundance. For example, one may have a data set consisting of customer purchasing records with attributes as stated as in Table 1. There are six numerical variables, three nominal variables and one ordinal variable:

Num1 – amount spend per visit;
Num2 – shopping time spent;
Num3 – age;
Num4 – number of visits within a month;
Num5 – number of purchases within a month;
Num6 – income;
Nom1 – type of purchasing products;
Nom2 – residential area;
Nom3 – service satisfaction level;
Ord1 –gender.

Suppose the company is going to launch a marketing campaign for promotion. Before making any promotion decision, very often it is desirable to classify its customers into two groups: *Loyal* or *Contingent*, see for instance [6]. Moreover, the company manager may intend to fit regression models (supervised learning) with amount spend per visit as the dependent variable for both types of customers in the hope of tackling different customer characteristics. In this situation, one cannot directly implement any regression model to the whole data set because it constitutes a vast mixture of information for both *Loyal* or *Contingent*. Thus, clustering procedure (unsupervised learning) [1, 2] is required.

ID	Variables									
	Numerical						Nominal			Ordinal
	Num1	Num2	Num3	Num4	Num5	Num6	Nom1	Nom2	Nom3	Ord1
1	35	15	52	2	2	NA	4	Far	3	M
2	102	20	31	4	3	18000	10	Med	6	M
3	40	5	23	5	5	NA	5	Far	7	F
4	32	45	42	6	4	24000	8	Med	6	F
5	8	82	65	10	2	3000	4	Near	4	F
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1000	20	36	28	3	1	18000	2	Near	5	F

Table 1: Variables and Customer Purchasing Records.

In this paper, we aim at developing a model namely *Semi-Supervised Regression Model* (SSRM) to fulfill the practical need. The basic idea of semi-supervised clustering [4, 19] is that the clustering algorithm is activated by a limited amount of supervisions. These

supervisions indeed are constraints established by the labeled data. The semi-supervised clustering algorithms use the labeled data (the constraints) to train the unlabeled data so as to improve the clustering accuracy.

The approach of *Semi-Supervised Learning* inspires us to develop our SSRM. We grasp the idea of tagging attributes with labeled data or not labeled data. Here we assume all the numerical attributes are labeled data. One of these numerical attributes is termed as the dependent variable (for example the amount spend per visit in the example), and other numerical attributes are independent variables. On the other hand, we assume that all the categorical attributes are unlabeled data. SSRM uses a clustering algorithm to cluster all unlabeled attributes. The data set is divided into several clusters, then SSRM computes a regression model for each cluster. The evaluation of model parameters is achieved by measuring a function which consists of a weighted sum of the least squares errors of these regression models and the dissimilarity measures among the categorical attributes. Such function values are used to train the model parameters and the partitioning of the data set. The above procedures can be repeated until the function values cannot be further improved (see Figure 1). In this paper, we will present experimental results on synthetic and real data sets to illustrate that the proposed method is effective.

The outline of this paper is as follows. In Section 2, we present the mathematical formulation of SSRM. In Section 3, experimental results are given to demonstrate the effectiveness of the proposed model. Finally, concluding remarks are given in Section 4.

2 The Formulation of SSRM

In this section, we present the notations which will be used throughout the discussion and the proposed SSRM.

2.1 Notations

In this paper, we have N (indexed by n) number of records. Attributes of these records are classified as either numerical or categorical. Numerical attributes include all those represented by real numbers and exist in a continuous space. A finite and discrete attribute is defined as categorical. We have R (indexed by r) numerical attributes and M (indexed by m) categorical attributes. Let

$$Z = (Z_1, Z_2, \ldots, Z_N)^T$$

be an $N \times R$ matrix for the numerical attributes,

$$C = (C_1, C_2, \ldots, C_N)^T$$

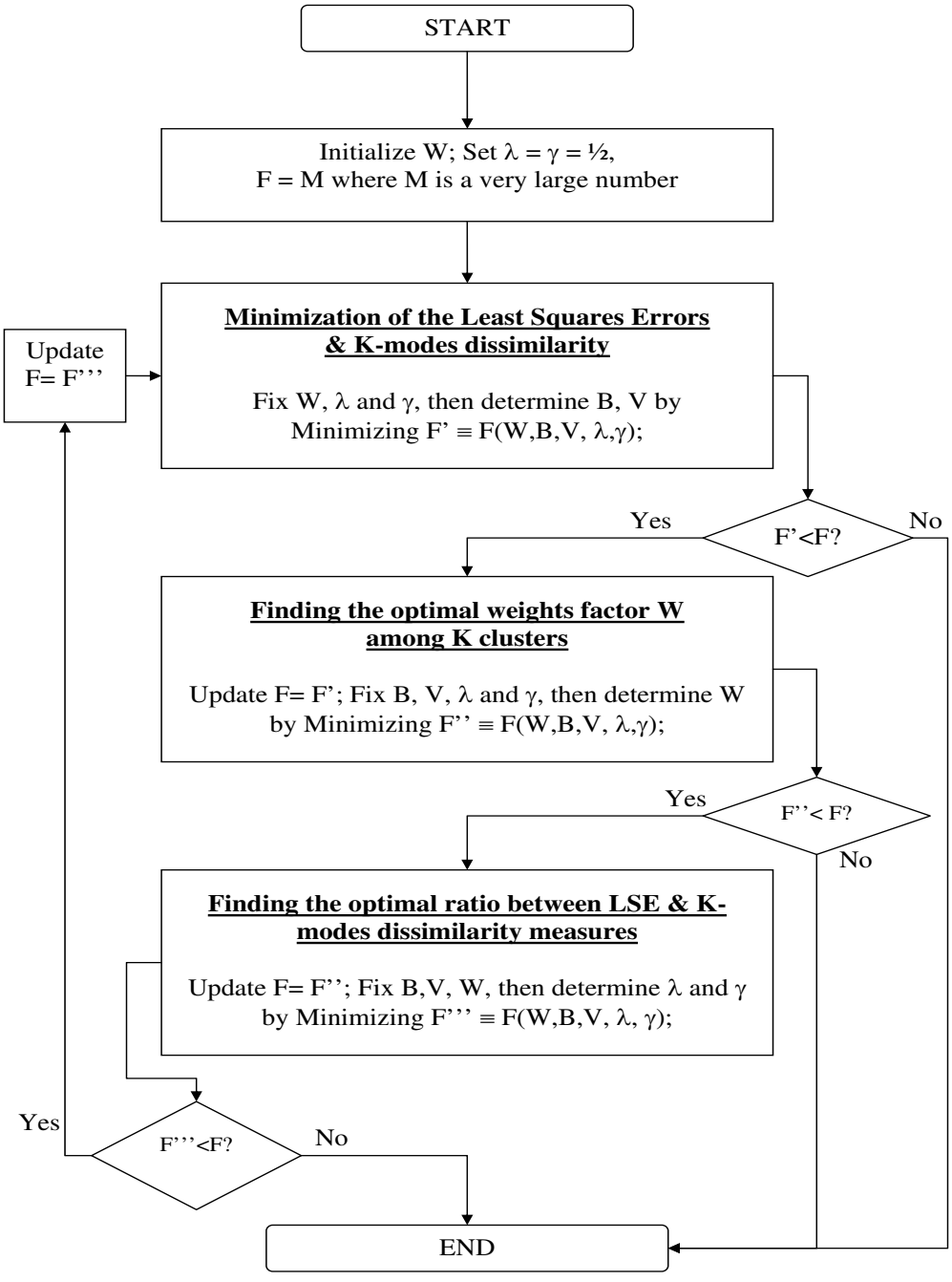

Figure 1: Flowchart of the SSRM algorithm.

be an $N \times M$ matrix for the categorical attributes, and

$$X = (Z, C) = (x_{nj}), \text{ where } n = 1, 2, \ldots, N; j = 1, 2, \ldots, R + M,$$

be an $N \times (R+M)$ matrix for all the records in which the first R columns are the numerical attributes, and the M categorical attributes as a consequence.

2.2 The Regression Model

The classical multivariate linear regression is effective in assessing the association among numerical variables. We have a single dependent variable Y and a collection of independent variables $Z_1, Z_2, \ldots, Z_L$ where $L = R - 1$. Each set of regression coefficients is labeled as $\beta_0^k, \beta_1^k, \ldots, \beta_L^k$, for $k = 1, 2, \ldots, K$ where k stands for the kth cluster. The multivariate linear regression model aims at predicting Y by using the linear predictor $B_k^T Z$, where

$$B_k = (\beta_0^k, \beta_1^k, \ldots, \beta_L^k)^T \quad \text{and} \quad Z = (1, Z_1, Z_2, \ldots, Z_L)^T.$$

For a given predictor of the above form, the error in the prediction of Y is the prediction error:

$$Y - B_k^T Z.$$

Since this error is random, it is customary to select B_k (for every cluster) to minimize the mean square error : $E(Y - B_k^T Z)^2$. For a finite set of data points, the error is measured by

$$\sum_{n=1}^{N} \left(Y^{(n)} - B_k^T Z^{(n)}\right)^2,$$

where

$$Y^{(n)} \quad \text{and} \quad Z^{(n)} = [1, Z_1^{(n)}, Z_2^{(n)}, \ldots, Z_L^{(n)}]$$

refer to the attributes of the nth record.

2.3 The Clustering Algorithm

As mentioned in Section 1, the main aim of our proposed algorithm is to partition the data set into K clusters. The clustering algorithm is the K-modes algorithm [7, 10, 12] which uses the K-means paradigm to cluster categorical data. Let $W = [w_{n,k}]$ be an N-by-K matrix represents the partitioning of all records into K clusters, where one record is dedicated to one cluster, and let

$$V = (V_1, V_2, \ldots, V_K)^T$$

be an K-by-M matrix representing a set of K-modes (that are centers) for those K clusters. The objective here is to find W and V that minimize the functional:

$$F(W,V) \quad = \quad \sum_{k=1}^{K}\sum_{n=1}^{N} w_{n,k} d(V_k, C_n), \tag{1}$$

subject to

$$w_{n,k} \in \{0,1\} \quad , 1 \leq k \leq K, \quad 1 \leq n \leq N, \tag{2}$$

$$\sum_{l=1}^{K} w_{n,k} = 1, \qquad\qquad 1 \leq n \leq N. \tag{3}$$

Here $K(\leq N)$ is the known number of clusters, and the simple matching dissimilarity $d(\cdot,\cdot)$ measures the distance between two records, which exists in vector form and is defined as follows:

$$d(X_i, X_l) = \sum_{r=R+1}^{M+R} \delta(x_{i,r}, x_{l,r}) \tag{4}$$

where

$$\delta(x_{i,r}, x_{l,r}) = \begin{cases} 0, & x_{i,r} = x_{l,r}, \\ 1, & x_{i,r} \neq x_{l,r}. \end{cases}$$

It is easy to verify that the function $d(\cdot,\cdot)$ defines a metric space on the set of categorical attributes.

The minimization of F in (1) with the constraints in (2) and (3) forms a class of constrained nonlinear optimization problems whose solution is unknown [3]. The usual method for the optimization of F in (1) is to use partial optimization for V and W. Here we employed Huang and Ng's K-modes algorithm [11]. They have shown that the K-modes algorithm converges in a finite number of iterations. For the K-modes algorithm, we refer to [7, 12] for details. Here we first fix V and find necessary conditions on W to minimize F. Then, we fix W and minimize F with respect to V. A frequency-based method for updating V, and calculation of W for a given V has been proposed by Huang [10].

2.4 The Semi-Supervised Regression Model

The clustering (unsupervised learning) and regression (supervised learning) models work extremely well for data in specific natures. In this subsection, we introduce a SSRM model that integrates these two algorithms to tackle both numerical and categorical data.

In the proposed SSRM, we estimate the best fit parameters by an iterative algorithm. There are two objectives in the proposed SSRM:

1. The minimization of the least squares errors for all sets of regression coefficients.

2. The minimization of the K-modes dissimilarity measures of all categorical attributes.

These two objectives are incorporated into a single objective function by introducing the weighting factors $\lambda_{\{\cdot,\cdot\}}$ and $\gamma_{\{\cdot,\cdot\}}$. Mathematically speaking, we have the following minimization problem:

$$\text{Min} \quad F(W,B,V,\lambda,\gamma) = \sum_{k=1}^{K}\sum_{n=1}^{N} w_{n,k}\left\{\lambda_{n,k}^{\eta}\{\left[Y^{(n)} - B_k^T Z^{(n)}\right]^2 + \gamma_{n,k}^{\eta} d(V_k, C_n)\right\} \tag{5}$$

subject to (2), (3) and

$$\begin{cases} \lambda_{n,k}, \gamma_{n,k} \geq 0, & 1 \geq k \geq K, \quad 1 \leq n \leq N, \\ \lambda_{n,k} + \gamma_{n,k} = 1, & 1 \leq k \leq K. \end{cases} \tag{6}$$

Here η is a control variable that found its best performance in the range $[1, \infty)$ from experimental experience. Since the minimization problem is highly non-linear, iteration method are employed to improve the estimation of the parameters. The iteration method may terminate at a local minimum, therefore different initial guesses have been tried in our numerical experiment. The SSRM algorithm reads:

The SSRM Algorithm.

Step 1. Choose an initial matrix W and set $\lambda_{n,k} = \gamma_{n,k} = \frac{1}{2}$, for all n and k

Step 2. Given W, λ and γ, determine B and V such that $F(W,B,V,\lambda,\gamma)$ is minimized. If the objective function value is improved, goto Step 3, otherwise stop.

Step 3. Given B, V, λ and γ, determine W such that $F(W,B,V,\lambda,\gamma)$ is minimized. If the objective function value is improved, goto Step 4, otherwise stop.

Step 4. Given B, V, W, determine λ and γ such that $F(W,B,V,\lambda,\gamma)$ is minimized. If the objective function value is improved, goto Step 2, otherwise stop.

We note that Step 2 can be determined by solving a least squares problem. Step 3 can be found by counting the number of dominant categorical attributes [10, 11]. For the W in Step 3, the minimizers are given by (see for instance [10])

$$w_{n,l} = \begin{cases} 1, & \text{if } \lambda_{n,l}^{\eta}\left[Y^{(n)} - B_l^T Z^{(n)}\right]^2 + \gamma_{n,l}^{\eta} d(V_l, C_n) \\ & \leq \lambda_{n,j}^{\eta}\left[Y^{(n)} - B_j^T Z^{(n)}\right]^2 + \gamma_{n,j}^{\eta} d(V_j, X_n), \quad \forall j \\ 0, & \text{otherwise.} \end{cases}$$

Data Sets	Number of of Numerical Attributes	Number of Categorical Attributes	Number of Clusters	Percentage of the Dominant Category
Sample 1	Two	Two	Two	60%
Sample 2	Two	Two	Two	40%
Sample 3	Two	Two	Three	60%
Sample 4	Two	Two	Three	40%

Table 2: Summary of data sets.

We remark that this minimum solution may not be unique, so $w_{n,l} = 1$ may arbitrarily be assigned to the first minimizing index l, and the remaining entries of this column are set to zero. Similarly, the other two parameters λ and γ can be found in Step 4. The flowchart summarizes the SSRM algorithm in Figure 1.

3 Experimental Results

In this section, we present experimental results of the proposed SSRM for both synthetic and real data sets.

3.1 Synthetic Data Sets

In order to test the performance of SSRM in varies settings, we conducted tests on several synthetic data sets with different characteristics in Table 2. For each categorical attribute, there are four categories.

For the numerical attributes, we used a variable σ to control the numeric variation from synthetic linear equations. The larger σ is, the higher variation of numerical attribute values will be. The generation of the numerical attribute values of such synthetic data sets is proposed in [9]. To obtain the average performance, we generated 100 test cases for each type of data sets.

Since the data sets contain real cluster labels for the data points, we use the external cluster validation method to evaluate the performance of the semi-supervised clustering algorithm in recovering the real clusters in the data. A clustering result is evaluated as follows. After a data set is clustered by the semi-supervised clustering algorithm, a new cluster variable is added to the data set to indicate the cluster each data point is assigned to. Using the cluster variable and the genuine class variable, we form a confusion matrix.

Let a_i be the maximal number of the points assigned to cluster i and whose genuine class is i. The clustering accuracy r is then defined as

$$r = \frac{1}{n} \sum_{i=1}^{k} a_i$$

where n is the number of records in the data set.

	Sample 1			Sample 2		
σ	Accuracy (%)	# of iteration	(λ, γ)	Accuracy (%)	# of iteration	(λ, γ)
0.1	99.00	49.27	(0.92,0.08)	95.93	48.50	(0.93,0.07)
0.3	96.30	24.97	(0.75,0.25)	92.03	25.24	(0.77,0.23)
0.5	92.68	19.74	(0.63,0.37)	86.34	19.77	(0.65,0.35)
0.8	88.35	20.27	(0.50,0.50)	78.06	18.90	(0.52,0.48)
1.0	85.12	46.38	(0.42,0.58)	76.81	19.34	(0.45,0.55)
1.5	80.72	46.38	(0.29,0.71)	72.06	25.79	(0.31,0.69)
2.0	68.14	111.5	(0.21,0.79)	70.33	37.89	(0.22,0.78)

Table 3: Average clustering results.

	Sample 3			Sample 4		
σ	Accuracy (%)	# of iteration	(λ, γ)	Accuracy (%)	# of iteration	(λ, γ)
0.5	80.22	24.85	(0.55,0.45)	73.87	29.65	(0.55,0.45)
1.0	71.02	33.84	(0.40,0.60)	60.74	42.44	(0.40,0.60)
1.5	61.27	53.29	(0.29,0.71)	53.29	57.77	(0.29,0.71)
2.0	56.47	71.09	(0.21,0.79)	48.76	67.61	(0.21,0.79)

Table 4: Average clustering results.

The average synthetic results and number of iterations required for convergence are summarized in Table 3 and Table 4. The results show that the proposed algorithm is efficient. In addition, we find the crucial factor that governs the accuracy of the SSRM is the noise standard deviation σ and the dominant percentage of the category value in the categorical attributes. Figure 2 shows the clusters of points and their corresponding regression plots. We observe from Figure 2 that the least-square estimates are very sensitive to outlying observation and this explains why the weighting parameter λ shifts from the numerical attributes to the categorical attributes ($\lambda \geq \gamma$) for a large σ. Because the disturbance from the numerical attributes becomes more significant, which in terms increase the difficulty to perform

regression analysis. The SSRM relies very much on the characteristics of the categorical attributes.

We also remark that the clustering accuracies of Sample 1 and Sample 3 are better than those of Sample 2 and Sample 4. This implies that the categorical attributes play an important role in the clustering process. We observe that there is a significant dividend on $\sigma = 2.0$ for the convergence rate. For σ smaller than this level, the SSRM is efficient. Figure 2 clearly shows the proposed algorithm perform quite well in the recovery of the underlying cluster structure and regression models.

Besides linear regression models, our current framework can also be applied to other nonlinear regression models. Here we consider quadratic regression models. Similar to the linear case, we constructed four types of synthetic data sets, namely Sample A, Sample B, Sample C and Sample D (see the generation of these data sets in [9]). For simplicity of discussion, we only test for quadratic equations in several settings. For each generated data set, we assigned two clusters and the dominant category involves in 60% of category values in each categorical attribute. We take the average of the results obtained by performing 100 tested cases. For the numerical attributes, the aim is to plot quadratic equations for each of those two clusters. For example, if there are two numerical attributes variables (stored in the $N \times 1$ vectors Z_1 and Z_2 where N is the number of records, and Z_2 corresponds to the dependent attributes) and two clusters are found in the data set, the regression models would be in the form:

$$Z_2^T = \alpha_\cdot^T [1, Z_1, Z_1^2]$$

where $\alpha_\cdot$ is a 3×1 vector with regression coefficients as entries. Here we try different initial values ti get the best solution. For Sample A, B and D, we considered initials

$$\alpha_1 = (-19, 2, 0) \quad \text{and} \quad \alpha_2 = (-31, 2, 0).$$

For Sample C, we considered initials

$$\alpha_1 = (9, 0, 0) \quad \text{and} \quad \alpha_2 = (7, 0, 0).$$

Experimental results are reported in Table 5. The results demonstrate that the performance of semi-supervised clustering models is quite well. High clustering accuracy results are obtained. Similarly to those of linear regressions. Figures 3 and 4 show the clusters of points and their corresponding regression plots. The figures clearly show that the proposed algorithm perform quite well in the recovery of the underlying cluster structure and regression models.

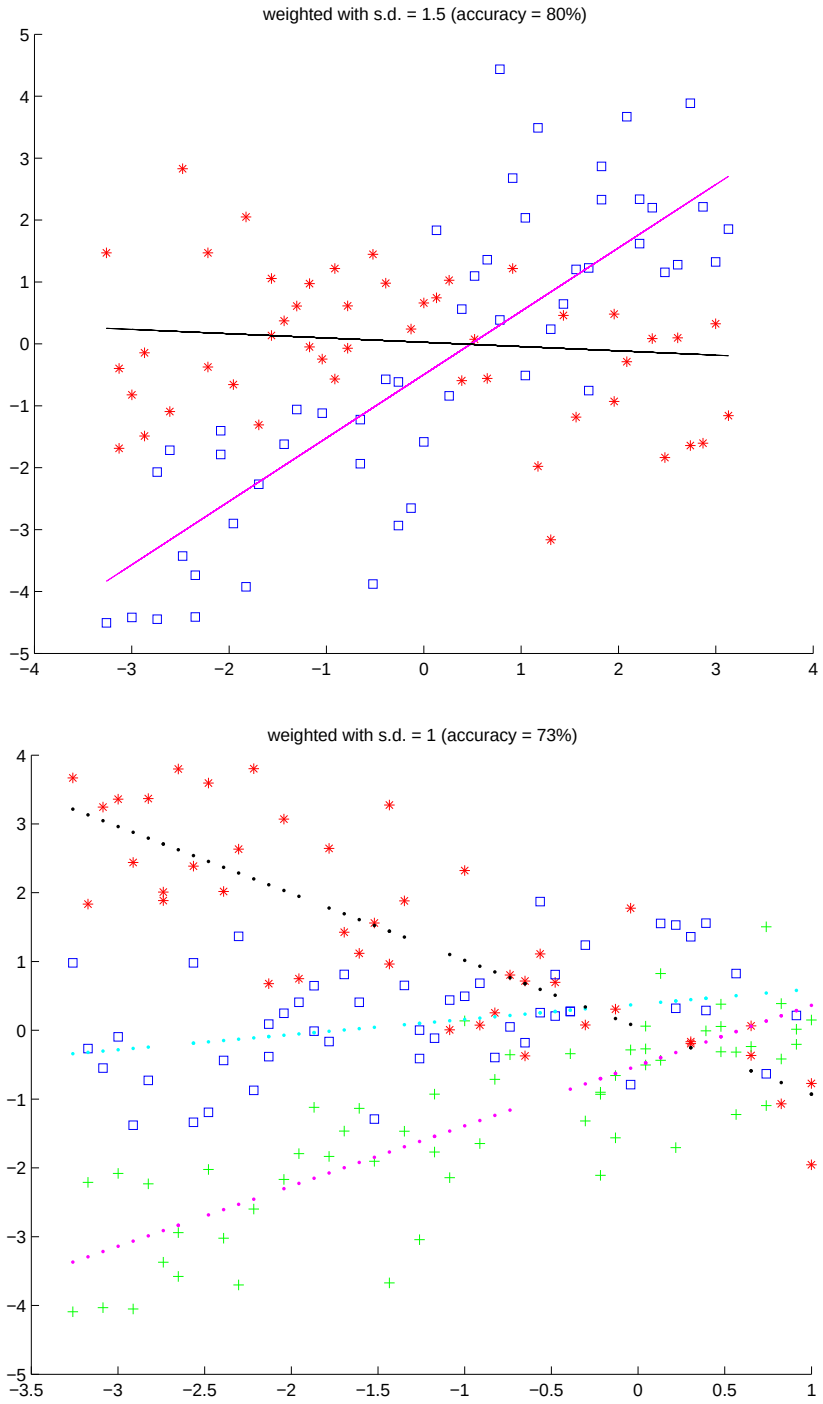

Figure 2: Plots of two regression models for Sample 1 (upper) and three regression models for Sample 3 (bottom) with $\sigma = 1.5$ and 1.0 respectively. ($\star$, $\square$, $+$) refer to different data set labels.

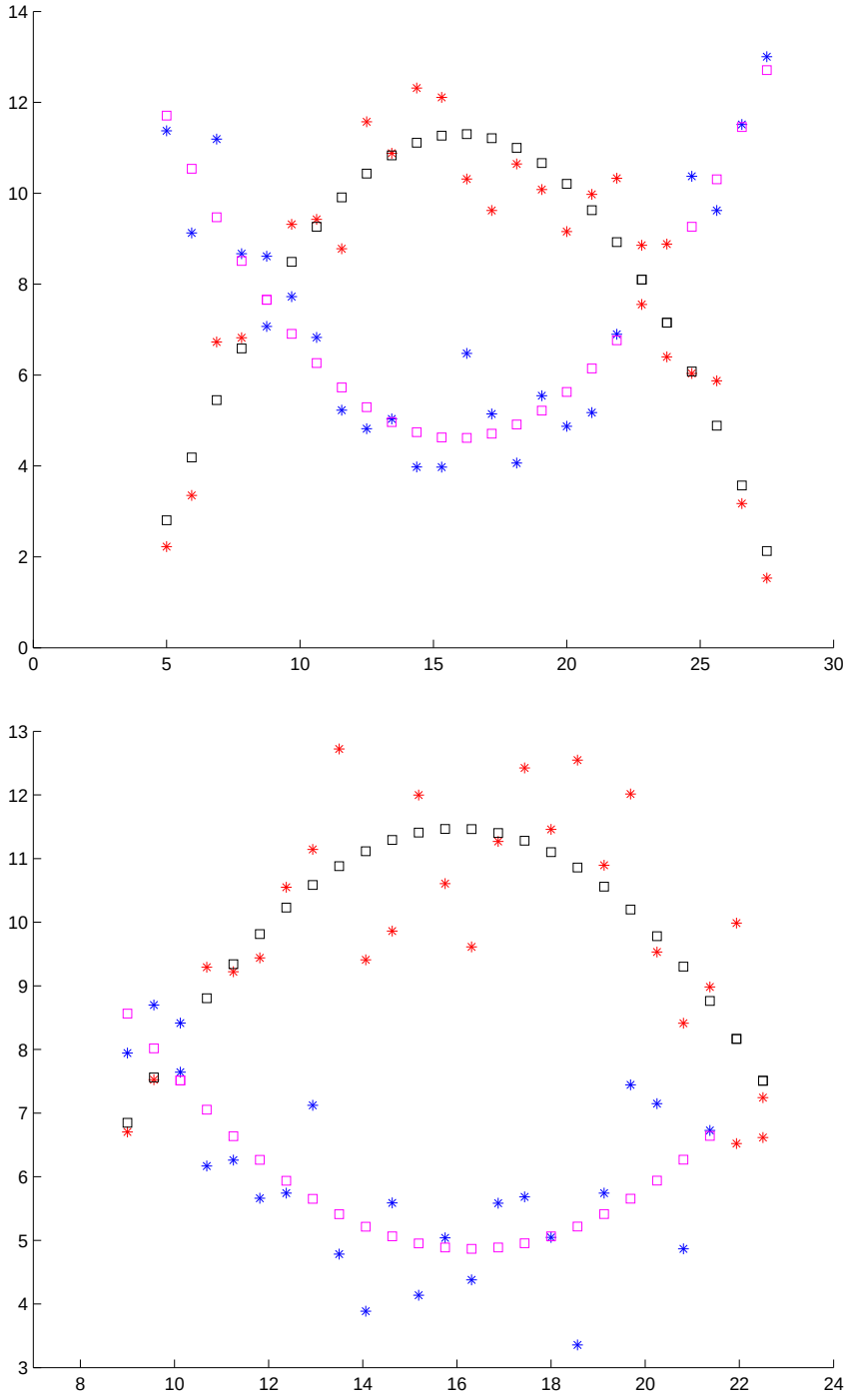

Figure 3: Plots of Sample A (upper) and Sample B (bottom), both with $\sigma = 1.5$.($\star$) refers to different data set labels.

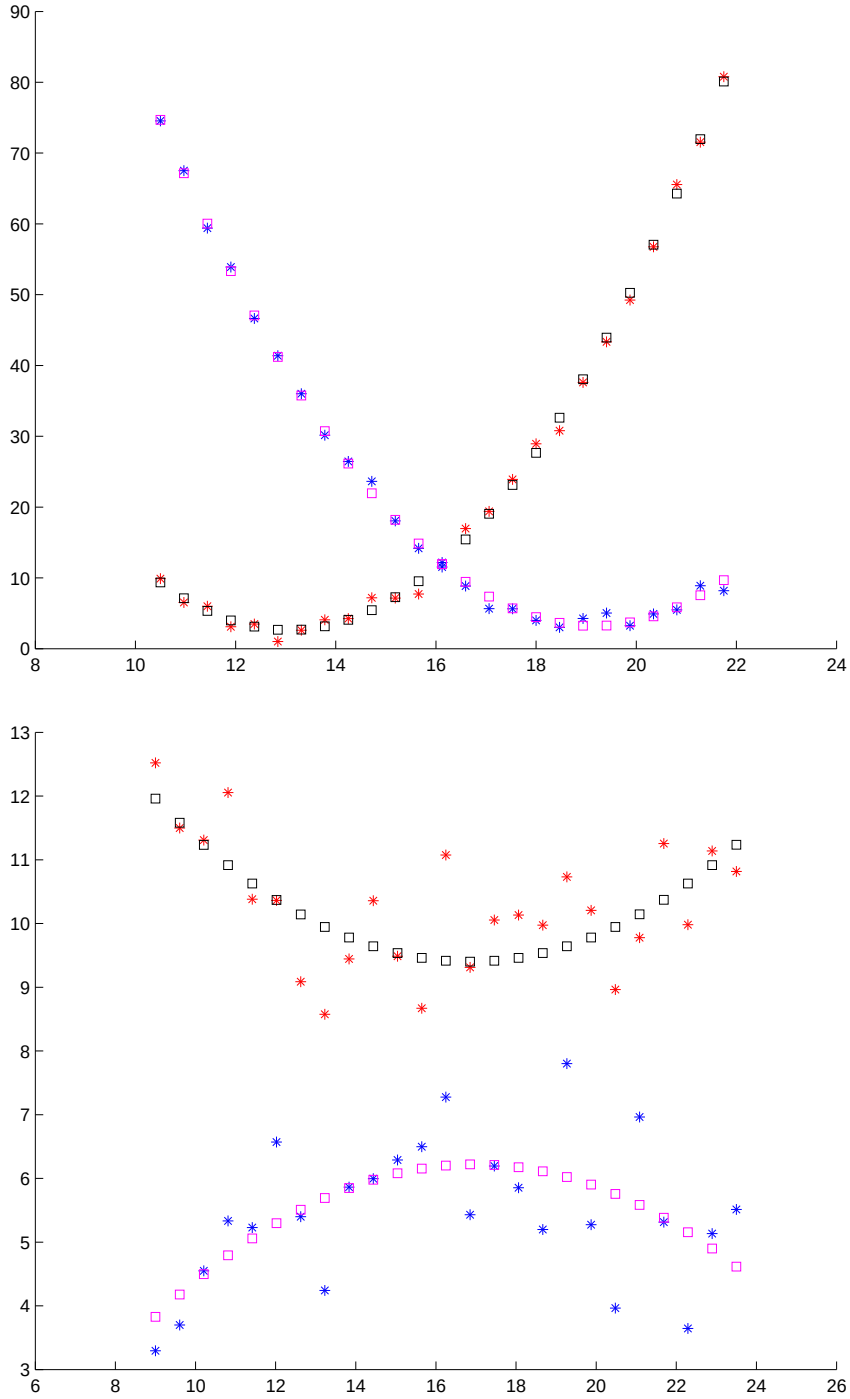

Figure 4: Plots of Sample C (upper) and Sample D (bottom), both with $\sigma = 1.5$. ($\star$) refers to different data set labels.

	Sample	$\sigma = 0.1$	$\sigma = 0.3$	$\sigma = 0.5$	$\sigma = 0.8$	$\sigma = 1.0$	$\sigma = 1.5$	$\sigma = 2.0$
Accuracy (%)	A	99.11	97.84	96.89	95.37	93.88	85.84	77.54
	B	98.94	97.58	96.58	94.88	93.06	89.22	85.14
	C	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	D	100.0	100.0	99.94	99.84	99.84	99.44	99.32
Number of Iterations	A	69.20	31.44	20.37	17.09	15.28	10.74	10.38
	B	67.17	30.15	19.06	15.67	13.94	11.72	10.67
	C	118.9	38.29	27.74	19.91	16.55	12.94	11.33
	D	228.1	130.1	92.78	67.89	67.89	42.83	34.73
(λ, γ)	A	(0.87, 0.13)	(0.61, 0.39)	(0.46, 0.54)	(0.33, 0.67)	(0.28, 0.72)	(0.14, 0.80)	(0.09, 0.91)
	B	(0.87, 0.13)	(0.62, 0.38)	(0.47, 0.53)	(0.33, 0.67)	(0.28, 0.72)	(0.20, 0.80)	(0.15, 0.85)
	C	(0.72, 0.28)	(0.47, 0.53)	(0.35, 0.65)	(0.25, 0.75)	(0.22, 0.78)	(0.16, 0.84)	(0.13, 0.87)
	D	(0.56, 0.44)	(0.25, 0.75)	(0.16, 0.84)	(0.09, 0.91)	(0.09, 0.91)	(0.05, 0.95)	(0.03, 0.97)

Table 5: Average clustering results for quadratic regression models.

3.2 A Real Data Set

To further evaluate the SSRM algorithm, we implemented it for a real data set – the German data set [22]. The German data set has 1000 consumer credit records. The records consist of both numerical and categorical attributes. For the credit records, the data set can be classified into two groups: Good Credit (on-time return loan) and Bad Credit (bad debt incurred). Thus, we can use this real classification to examine the SSRM clustering performance. All the records were studied as there is no missing value in the data set. The categorical data are: *Status of existing checking account, Credit history, Purpose, Savings account/bonds, Present employment, Personal status and sex, Other debtors / guaranteers, Property, Other installment plans, Housing, Job, Telephone*, and *foreign worker*. The average clustering accuracy for our model, compared with the given Good/Bad clusters, is 68.40%.

There are some obvious differences between the two clusters. For the *Telephone* attributes, most Bad debt customers registered their telephone numbers under their customer records whereas those Good debtors didn't. Moreover, for the *Present employment*, most Good debtors have employed in the present position for less than four years. On the contrary, a large number of Bad debtors have worked at a stable position which last for more than seven years. This is indeed a counter-pole for a common conception that a stable worker would be capable or would take up the responsible for its own debt! There are eleven categorical choices for *Purpose*. For Good debtors, if the loan is for purchase, they use the money to buy day-to-day necessities like television or radio, whereas most of the Bad debtors tend to buy luxury goods like car.

The numerical attributes are: *Duration in month* (z_1), *Installment rate in percentage of disposable income* (z_2), *Present residence since* (z_3), *Age in years* (z_4), *Number of existing*

credits at this bank (z_5), *Number of people being liable to provide maintenance for* (z_6), and *Credit amount* (z_7). The regression models obtained from SSRM are given as follow:

$$z_7 = 50.8 + 1564.4z_1 + 111.4z_2 - 588.8z_3 - 14.7z_4 + 8.6z_5 + 98.8z_6 \qquad \text{(Good Credit)}$$

$$z_7 = -659.3 + 4558.7z_1 + 232.3z_2 - 1214.3z_3 + 146.1z_4 + 26.1z_5 - 498.6z_6 \quad \text{(Bad Credit)}$$

Differences and similarities are found. The positive intercept indicates that for a Good debtor with $z_j = 0$ for $j = 1, 2, 3, 4, 5, 6$, the Credit amount would be \$50.8. This would be interpreted as, in the absence of all regressors, Good debtors have the tendency to acquire more money. However, for the Bad debtors, the initialization for borrowing is negative. Indeed, these intercepts would evaluate the effect brought by other factors, like psychological differences, that have not been studied in this research. The regressor *Age* performs differently in these two clusters. For Good Credit, an older debtors tend to borrow less. However, the regression coefficient for z_4 indicates that a unit increase of Bad debtor's age will raise the Credit amount by \$146.1. Also, Credit amount for Bad Credit is negatively related to the number of people being liable to provide maintenance (z_6). But two variables are positively related for Good Debtors. This is reasonable because the debtors with good credit record are more likely to find guaranteer. Although credit behavior is not the main objective of our study, our SSRM works well in this situation.

4 Concluding Remarks

In this paper, we proposed a combined mathematical model, SSRM, for modeling both numerical and categorical data. The cornerstone for this SSRM is integrating two conventional mathematical methods, K-mode clustering and regression model. A major advantage is its flexibility. SSRM can deal with both numerical and categorical data for a wide variety of data. This model is particularly useful in analyzing data with a slight dominant categorical data. It is also remarkable to notice the fast convergence rate of SSRM as demonstrated by our experimental results.

References

[1] Anderberg, M. R. (1973). *Cluster analysis for applications*. New York: Academic.

[2] Ball, G. H. and Ball, D. J. (1967). A clustering technique for summarizing multivariate data, *Behavioral Science*, v.12, pp. 153-155.

[3] Bezdek, J. C. (1980). A convergence theorem for the fuzzy ISODATA clustering algorithms, *IEEE Trans. Pattern Anal. Machine Intell.*, v. PAMI-2, pp. 1-8.

[4] Blum, A. and Mitchell, T. (1998). Combining labeled and unlabeled data with co-training, *Proceedings of the 11th Annual Conference on Computational Learning Theory*, pp.39-48.

[5] Chan, E. Y., Ching, W.K., Ng, M. K and Huang, J. Z. (2004). An optimization algorithm for clustering using weighted dissimilarity measures, *Pattern Recognition*, v.37, pp. 943-952.

[6] Ching, W. K., Ng, M and Wong, K. K. (2004). Hidden Markov model and its applications in customer relationship management, *IMA Journal of Management Mathematics*, v.15, pp. 13-24.

[7] Chaturvedi, A., Green, P. and Carroll, J. (2001). K-modes clustering, *Journal of Classification*, v18, pp. 35-55.

[8] Duda, R. O. and Hart, P. E. (1973). *Pattern classification and scene analysis*, New York: John Eiley & Sons.

[9] Hathaway, R. J. and Bezdek, J. C. (1993). Switching regression models and fuzzy clustering, *IEEE Transactions on Fuzzy Systems*, v. 1, pp. 195-204.

[10] Huang, J. Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values, *Data Mining Knowledge Discovery*, v. 2, pp. 283-304.

[11] Huang, J. Z. and Ng, M. K. (1999). A fuzzy k-modes algorithm for clustering categorical data, *IEEE Trans. on Fuzzy Systems*, v. 7, pp. 446-452.

[12] Huang, J. Z. and Ng, M. K. (2003). A note on K-modes clustering, *Journal of Classification*, v. 20, pp. 257-261.

[13] Hubert, M. and Rousseeuw, P. J. (1997). Robust regression with both continuous and binary regressors, *Journal of Statistical Planning and Inference*, v. 57, pp. 153-163.

[14] Jain, A. K. and Dubes, R. C. (1988). *Algorithms for clustering Data*, Englewood Cliffs, NJ: Prentice-Hall.

[15] Kleinbaum, D. G., Kupper, L. L., and Muller, K. E. (1998). *Applied regression analysis and other multivariate methods*, Pacific grove, California: Duxbury Press.

[16] McLachlan, G. J. (1992). *Discriminant analysis and statistical pattern recognition*, Wiley, New York.

[17] McCullagh P. (1980). Regression models for ordinal data, *Journal of the Royal Statistical Society-Series B*, v. 42, pp. 109-142.

[18] Menard, S. W. (2001). *Applied logistic regression analysis*, Sage Publications.

[19] Nigam, K., McCallum, A.K., Thrun, S. and Mitchell, T. (2000). Text classification from labeled and unlabeled documents using EM. *Machine Learning*, v. 39, pp.103-134.

[20] Stahel, W. A. (2004). Purposes and strategies in regression analysis, *Journal of Statistical Planning and Inference*, v. 122, pp.175-186.

[21] Tabachnick, B. G. and Fidell, L. S. (2001). *Using multivariate statistics*, Boston, 4th ed.

[22] WWW URL http://www.liacc.up.pt/ML/statlog/datasets/german/german.doc.html

[23] Wedel, M. and Kamakura, W. A. (2000). textslMarket segmentation : conceptual and methodological foundations, Kluwer Academic. 2nd ed., Boston.