Iranian Journal of Operations Research
Vol. 4, No. 1, 2013, pp. 55-74

Multi-Group Classification Using Interval Linear Programming

B. Izadi[*,1], B. Ranjbarian[2], S. Ketabi[3], F. Nassiri-Mofakham[4]

Among various statistical and data mining discriminant analysis proposed so far for group classification, linear programming discriminant analysis has recently attracted the researchers' interest. This study evaluates multi-group discriminant linear programming (MDLP) for classification problems against well-known methods such as neural networks and support vector machine. MDLP is less complicated as compared to other methods and does not suffer from having local optima. This study also proposes a fuzzy Delphi method to select and gather the required data, when databases suffer from deficient data. In addition, to absorb the uncertainty infused to collecting data, interval MDLP (IMDLP) is developed. The results show that the performance of MDLP and specially IMDLP is better than conventional classification methods with respect to correct classification, at least for small and medium-size datasets.

***Keywords**: Multi-group interval linear programming, Classification problem, Fuzzy Delphi feature selection.*

Manuscript received on 25/1/2013 and accepted for publication on 6/6/2013.

1. Introduction

The applications of classification methods are wide-ranging and the advent of powerful information systems since the mid-1980s has renewed interest in classification techniques (Pai et al. [21]). Differentiating among patients with strong prospects for recovery and those highly at risk, among customers with good credit risks and poor ones, or between promising new firms and those likely to fail, are among the most known applications (Youssef and Rebai [26]). Managers specially use classification techniques to make decisions in different business operation areas.

At its broadest, classification could cover any context in which some decision or forecast is made on the basis of currently available information. A classification procedure is then a formal method for repeatedly making such judgments in new situations (Michie and Spiegelhalter [18]). For instance, rather than targeting all customers equally or providing the same incentive offers to all customers, managers can select those customers who meet some profitability criteria based on purchasing behaviors (Dyche and Dych [5]). However, due to the nature of classification problem, a spectrum of techniques is needed because no single technique always outperforms others under all situations (Johnson and Wichern [10]). Various methods have been proposed for solving

* Corresponding Author.
[1] Department of Management, Faculty of Administrative Sciences and Economics, University of Isfahan, Iran.
E-mail: izady.bahram@gmail.com
[2] Department of Management, Faculty of Administrative Sciences and Economics, University of Isfahan, Iran.
[3] Department of Management, Faculty of Administrative Sciences and Economics, University of Isfahan, Iran.
[4] Department of Information Technology Engineering, Faculty of Engineering, University of Isfahan, Iran.

classification problems which can be divided into two categories: parametric and non-parametric discriminant methods. There are no pre-defined assumptions in non-parametric methods. However, parametric methods make strong parametric assumptions such as multivariate normal populations with the same variance/covariance structure, absence of multi co-linearity, and absence of specification errors (Meyers et al. [17]). Classification methods can also be grouped as statistical approaches such as Linear Discriminant Analysis (LDA) and Logistic Regression (LR), Artificial Intelligence (AI) techniques such as Artificial Neural Network (ANN) and operation research techniques such as Linear Programming (LP) and Goal Programming (GP).

The earliest linear discriminant method was proposed by Fisher in 1936. This linear method of discrimination requires the sample to be distributed normally and the variance-covariance matrices of the two groups to be homogeneous. Mangasarian [15] also was the first who used LP methods for classification problems. Linear programming methods have some advantages over other approaches which can be summarized as follows. First, there is no assumption about the functional form and hence it is distribution free. Second, they are less sensitive to outliers. Third, they do not need large datasets. Nonetheless, linear programming methods also have a disadvantage, which is the need for lengthy computation. However, the immense increase in computing power and drop in computing cost have over shadowed the disadvantage and made the LP methods practical.

Our work here makes use of Linear Discriminant Analysis, Logistic Regression, Support Vector Machine, and Artificial Neural Network vis-a-vis Multi-Group Discriminant Linear Programming (MDLP) discriminant analysis to classify the customers of an Internet Service Provider company based on their real demographic data including age, gender, education, income, and purpose, in order to show the strength and accuracies of classification models, specially for MDLP.

One major problem in classification methods, specially for small businesses or E-businesses, is the dispersed, deficient and redundant data in their databases. For this reason, researchers usually use synthetic or simulated data; however, it is not realistic. That is why we propose the application of fuzzy Delphi method for the selection of the required features. This helps companies to classify their customers with their current rudimentary databases; otherwise, they should wait for months or years for depositing the data and it is obviously undesirable in today's competitive environment. However, collecting some kind of sensitive customer data, such as income, using the proposed Delphi process, embrace some uncertainty due to the biases resulting from different sources like questions, respondent, interviewer and interview situation (Ziniel [27]). Furthermore, businesses usually use interval tables for collecting sensitive data during online customer registration to decrease their resistance for disclosing their personal data. Considering this situation, we develop MDLP method in the interval form (IMDLP) to absorb uncertainties imposed on customer data which eventually affect the classified boundaries of the collected data.

To achieve our aim, the ISP company dataset was used to segment the customers based on an RFM (Recency, Frequency and Monetary) model by different clustering methods such as K-means, Self-Organizing-Map, Two-Step Clustering and Data Envelopment Analysis. The customers are grouped in three distinct segments. Each segment comprises its own average R, F and M value giving an indication of customer lifetime values, (CLV). The acquired information obtained in clustering step is utilized for our purposes.

The organization of this paper is as follows. A brief description of various classification methods is given Section 2. Section 3 presents data preparation and Section 4 applies fuzzy Delphi. Section 5 shows the computational results and exhibits the performance of different classification methods including MDLP and IMDLP. We conclude in Section 6.

2. Classification Methods

Well-known classification methods including Logistic Regression, Linear Discriminant Analysis, Artificial Neural Network, Support Vector Machine, Multi-group Discriminant Linear Programming (MDLP) and the proposed Interval MDLP (IMDLP) are described here.

2.1. Logistic Regression (LR)

Logistic regression is a modeling procedure where a set of independent variables are used to model a dichotomous criterion variable using maximum likelihood estimation (MLE) procedure. Therefore, it is appropriate for direct marketers who would like to model a dichotomous variable such as presence/absence, success/failure, buy/don't buy, default/don't default, and survive/die. The availability of sophisticated statistical software and high speed computers has further increased the utility of logistic regression. The predicted value is the response probability, which varies from zero to one (McCarty and Hastak [16]). In logistic regression, the probability of a dichotomous outcome is related to a set of predictor variables in the form of the following function:

$$\ln\left(\frac{p}{1-p}\right) = \beta X = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n, \tag{1}$$

where p is the probability of the outcome of interest, β_0 is the intercept term, and β_i ($i=1,\ldots,n$) represents the coefficient associated with the corresponding explanatory variable x_i , $i=1,\ldots,n;$ see Shmueli [22]. The dependent variable is the logarithm of the *odds*, [*log* [p /($1-p$)]], which is the logarithm of the ratio of two probabilities of the outcome of interest. For instance, if we are interested to measure the effect of independent variables such as consumption of tobacco cigarette and alcohol on blood status indices and 25 of 100 persons have indications in their blood, then it can be said the odds is 25 to 75 or one-third. However, when the probability of occurrence is too high or too low, odds go to infinity. Then, the logarithm of odds is usually used to resolve this problem. The logarithm of odds is named logit. If the logit is negative, then it means the odds are against the event occurrence and vice versa. If the odds are 50-50, then the logit is zero. Here, we use multinomial logistic regression, because the dependent variable is a three-class categorical variable which indicates three pre-defined market segments.

2.2. Fisher Linear discriminant Function (LDF)

Database marketers frequently use discriminant analysis as an alternative to logistic regression. Discriminant analysis is a multivariate technique identifying variables which explain the differences among several groups and classify new observations or customers into the previously defined groups (Blattberg et al. [2]). This method separates classes by linear frontiers to group the data to be classified around the center of gravity (average) of each class and to create a linear hyper plane corresponding to the classes. This method requires certain assumptions: the normality distribution

of the samples, and homogeneity of the variance-covariance matrices (Youssef and Rebai [26]). If we have *n* variables, then the discriminant function is:

$$Z_i=b_0+b_1X_{1i}+b_2X_{2i}+ \ldots +b_nX_{ni}, \quad (2)$$

where X_{ij} is *i*th individual's value of the *j*th independent variable, b_j is discriminant coefficient for the *j*th variable, Z_i is *i*th individual's discriminant score and Z_c is critical value for the discriminant score. If $Z_i>Z_c$, then individual *i* belongs to class 1, otherwise to class 2. If there are two classes, then the separating boundary is a straight line. Generally, the classification boundary is an (*n-1*) dimensional hyperplane in *n* dimension space (Morrison [19]). LDF depends on parameters b_0, $b_1,\ldots,b_n$ and an algorithm of training is used to determine these parameters. This algorithm aims to satisfy the criterion associated with the model which generally aims to minimize the error of classification.

2.3. Artificial Neural Network (ANN)

Artificial neural network models are fascinating because they are based on the intuitive concept of imitating the structure of neurons that constitute the human brain, because ANN learns and generalizes from the external inputs based on a model in analogy with human brain. There are different architectures or topologies for ANN such as single-layer, multi-layer, feed-forward and recurrent. Feed-forward multi-layer is also called multi-layer perceptron (MLP) which is preferred due to its simple architecture and its practical success in solving approximation problems. An MLP network consists of a set of source nodes forming the input layer, one or more hidden layers of computing nodes, and an output layer which refers to the desired output of the system (Celebi and Bayraktar [3]). Neurons in the input layer correspond to the base evaluation criteria which are grouped under target performance measure. Here, there are totally thirteen input neurons. The neurons are five demographic variables of customers with different levels. i.e. age, gender (0=female, 1=male), education (Below Diploma=0, Diploma=1, University Diploma=2, B.S.=3, M.S.=4, PhD.=5), income, purpose (1= doing research, 2= doing business, 3= entertainment and social networks).

There are different training methods for building neural network models. The Radial Basis Function Network (RBFN) uses a technique to partition the data based on values of the target field. A pruning method starts with a large network and removes the weakest units in the hidden and input layers as training proceeds. In exhaustive pruning method, network training parameters are chosen to ensure a very thorough search of the space of possible models to find the best one. Dynamic training method modifies the topology by adding or removing hidden units as training progresses [28]. The data set is usually divided into two subsets, one being used for training the ANN and the other being used for validation (Blattberg et al. [2]).

2.4. Support Vector Machine

The foundations of Support Vector Machines (SVM) have been developed by Vapnik [24] and gained popularity due to many promising features such as adequate empirical performance. The goal of support vector machine is to find the particular hyperplane (called the optimal hyperplane) which maximally separates two classes (Flach [6]). The observations closest to the optimal hyperplane are called the support vectors X_s. The optimal weighting vector can be derived by solving a quadratic optimization problem. Support vectors can be derived once the optimal weighting vector is determined. These vectors play a critical role in support vector machines. Since support vectors lie closest to the decision surface, they are the most difficult observations to classify. However, only these closest instances are required to classify new instances so that the remaining instances play no role in predicting the class of new ones. It means that a set of support vectors uniquely defines the optimal hyperplane. However, the biggest disadvantage of the linear hyperplane is that it can only represent linear boundaries between classes (Witten and Frank [25]). One way of overcoming this restriction is to transform the instance space into a new "feature" space using a nonlinear mapping. A straight line in feature space does not look straight in the original instance space. That means that, a linear model constructed in feature space can represent a nonlinear boundary in the original space. Briefly speaking, the idea of support vector machines is based on two mathematical operations: (1) nonlinear mapping of the original instance space into a high dimensional feature space, and (2) construction of an optimal hyperplane to separate the classes. In other words, we need to derive the optimal hyperplane defined as a linear function of vectors drawn from the feature space rather than the original instance space (Blattberg et al. [2]).

2.5. Multi-group Discriminant Linear Programming (MDLP)

Linear Programming (LP) or, at its broadest, Mathematical Programming (MP) approaches are nonparametric and have attracted being many researchers interest, because these methods do not make strict assumptions about the data analyzed, are less influenced by outlier observations and are flexible in incorporating restrictions. The publication of the original LP models for the two-class classification by Freed and Glover [7] inspired a series of studies. Some of these studies reported pathologies of the earlier MP models, some provided diagnoses, and others offered remedies (Sun [23]). The method uses a weighting outline to establish a critical value or cutoff point that serve as a breakpoint between two successful and unsuccessful groups. Afterwards, Freed and Glover [7] proposed a set of interrelated goal programming formulations. They showed the potential of these formulations with the help of a simple example of assigning credit applicants to risk classifications. However, for decades, MP approaches have been limited to two-class methods because of the lack of powerful and simple models. Among the various MP models for classification problems, the model proposed by Lam et al. [13] have recently attracted researchers' interest (Sun [23]). They modified the earlier model proposed by Freed and Glover [7] for multi-group classification problem and proved that it is more powerful in terms of hit-rate criterion and its stability than statistical methods. Their model is as described next.

Suppose there are totally n observations distributed in m groups so that $n=n_1+n_2+\ldots+n_m$, where n_k is the number of observations in group k (G_k). If x_{ij} is value of the jth variable (attribute) for the

*i*th observation in the sample and we consider q variables, then for each pair of groups (u, v), u=1,...,m-1 and v=u+1,...,m, the minimization of the sum of deviations model is as follows:

$$\begin{aligned}
&\min \sum_{i \in G_u \cup G_v}^{n} d_i \\
&S.t. \\
&\sum_{j=1}^{q} w_j (x_{ij} - \bar{x}_j(u)) + d_i \geq 0, \quad \forall i \in G_u \\
&\sum_{j=1}^{q} w_j (x_{ij} - \bar{x}_j(v)) - d_i \leq 0, \quad \forall i \in G_v \\
&\sum_{j=1}^{q} w_j (\bar{x}_j(u) - \bar{x}_j(k)) \geq 1, \\
&d_i \geq 0, \forall i \in G_v v G_v,
\end{aligned} \tag{3}$$

where d_i is the deviation of the individual observations from cut-off scores (c), w_j is the weight of variable j and $\bar{x}_j(k)$ is the mean of the jth variable in group k (k=1, 2 ,..., m), defined as follows:

$$\bar{x}_j(k) = \sum_{i \in G_k} x_{ij} / n_k . \tag{4}$$

The objective function is to minimize the sum of all the deviations. The first two constraints force the classification scores of the observations in G_k to be as close to the mean classification score of group k as possible by minimizing d_i and the last constraint is a normalization constraint in order to prevent trivial values for discriminant weights.

Now, the calculated w_j values are used to obtain the values of classification scores in each group G_k:

$$S_i = \sum_{j=1}^{q} w_j x_{ij} \ , i \in G_k . \tag{5}$$

Then, the cut-off scores (C_{uv}), which indicate the separating boundary between the groups, are calculated by the following LP model:

$$\begin{aligned}
&\min \sum_{u=1}^{m-1} \sum_{v=u+1}^{m} \left(\sum_{i \in G_u} d_{iuv} + \sum_{i \in G_v} d_{iuv} \right) \\
&S.t. \\
&S_{iuv} + d_{iuv} \geq C_{uv} \ , \text{for } u = 1,...,m-1, \ \ v = u+1,...,m, \ \ i \in G_u \\
&S_{iuv} - d_{iuv} \leq C_{uv} \ , \text{for } u = 1,...,m-1, \ \ v = u+1,...,m, \ \ i \in G_v ,
\end{aligned} \tag{6}$$

where the C_{uv} are unrestricted in sign. This process converts the classification problem to $m(m\text{-}1)/2$ distinct two-group problems which is solved separately.

Another extension of this model is offered by Pai et al. [21] based on the model of Lam et al. [13] in which, the mean is substituted by median. They proved that this substitution increases the efficiency of the model with respect to hit rate, specially when the distribution of dataset is abnormal or skewed. Therefore, this study uses the median instead of mean in equation (3).

2.6. Interval Multi-group Linear Programming

It is pointed out that collecting sensitive customer data, by the Delphi method or in the form of interval table in online customer registration, incurs biases which, in turn, cause uncertainties to the harvested data, eventually affecting the classified boundaries of collected data. In other words, classified boundaries of collected data are not strict, specially for some variables like income. In the end, these uncertainties affect feasible region and classification results. To absorb these biases, it is proposed to modify model (6) in the form of interval MDLP. Model (7) below shows a matrix form of an interval linear programming problem:

$$\begin{aligned} &\min\ \mathrm{C}x \\ &s.t. \\ &Ax \geq b \\ &b \in [\underline{b}, \bar{b}]. \end{aligned} \tag{7}$$

We can rewrite as (8) below:

$$\begin{aligned} &\min\ Cx \\ &s.t. \\ &Ax = b \\ &\underline{b} \leq b \leq \bar{b}. \end{aligned} \tag{8}$$

Model (6) is now re-written as follows:

$$\begin{aligned} &\min \sum_{u=1}^{m-1} \sum_{v=u+1}^{m} \left(\sum_{i \in G_u} d_{iuv} + \sum_{i \in G_v} d_{iuv} \right) \\ &\mathrm{s.t.} \\ &d_{iuv} - C_{uv} \geq -S_{iuv}\ , \text{for } u = 1,...,m-1,\ \ v = u+1,...,m,\ \ i \in G_u \\ &d_{iuv} + C_{uv} \geq S_{iuv}\ , \text{for } u = 1,...,m-1,\ \ v = u+1,...,m,\ \ i \in G_v \\ &d_{iuv} \geq 0. \end{aligned} \tag{9}$$

Considering models (7) and (8), it is possible to write mode (9) as follows

$$\begin{aligned} &\min \sum_{u=1}^{m-1} \sum_{v=u+1}^{m} \left(\sum_{i \in G_u} d_{iuv} + \sum_{i \in G_v} d_{iuv} \right) \\ &\mathrm{s.t.} \\ &d_{iuv} - C_{uv} = -b_{iuv}\ , \text{for } u = 1,...,m-1,\ \ v = u+1,...,m,\ \ i \in G_u \\ &d_{iuv} + C_{uv} = b_{iuv}\ , \text{for } u = 1,...,m-1,\ \ v = u+1,...,m,\ \ i \in G_v \\ &\underline{S}_{iuv} \leq b_{iuv} \leq \bar{S}_{iuv}, \text{for } u = 1,...,m-1,\ \ v = u+1,...,m,\ \ i \in G_u, i \in G_v \\ &d_{iuv} \geq 0. \end{aligned} \tag{10}$$

Therefore, two different values, best and worst cases for lower and upper bounds, are obtained for classification scores. It is also possible to rely on the average of both.

3. Data Preparation

To show the efficiency of the aforementioned classification models in real settings, we use a dataset related to 6000 customers provided by Irangate Internet Service Provider (ISP) Company for a five years period from 2007 to 2012. Irangate is a major ISP company in Iran whose mission is to satisfy customers' needs for different kinds of internet services such as E-commerce and DSL. This company offers its customers different packages of bandwidth, time and charge of internet services.

The required data for classification purposes as scattered within databases of different company departments have been gathered, combined, refined and prepared. After removing redundant and inconsistent data, 5271 records have remained. Recency, Frequency and Monetary (RFM) model has been applied to segment the customers using clustering ensemble method in which K-means, Self-Organizing Map (SOM) and Two-Step clustering have been used. Cluster ensembles can be formed in a number of different ways such as the use of a number of different clustering techniques, the use of a single technique many times with different initial conditions and the use of different partial subsets of features or patterns (Kotsiantis and Pintelas [12]).
The K-means algorithm for partitioning is base on the mean value of the objects in the cluster; the term K-means was suggested by MacQueen [14] for describing an algorithm that assigns each item to the cluster with the nearest centroid or mean. The SOM is an unsupervised neural network learning algorithm and forms a mapping of the high-dimensional data to two-dimensional space and forms clusters to represent groups of nodes with similar properties (Kiang [11]). In two-step clustering, in the first step, cases are assigned to pre-clusters and in the second step, the pre-clusters are clustered using the hierarchical clustering algorithm.

Based on the clustering ensemble concept and using the aforementioned clustering techniques, three customer segments were identified with different internet usage pattern based on RFM model as shown in Table 1. RFM is a useful marketing technique to improve customer segmentation. It is used for analyzing customer behavior such as how recently a customer has purchased (recency), how often the customer purchases (frequency), and how much the customer spends (monetary) (Birant [1]).

Table 1. Three customer segments identified based on RFM model

	Count	Mean of Recency Score (Months)	Mean of Frequency Score (Times)	Mean of Monetary Score (Giga Bytes)
Cluster 1	2387	4.838	2.062	2.582
Cluster 2	1776	6.808	7.068	6.306
Cluster 3	1108	2.407	1.977	7.836

According to Table 1, the average recency, frequency and monetary of the segments are totally different. As shown, cluster 2 is the segment of high-valued customers, because the average recency, frequency and monetary of this segment is the highest. Cluster 3 also needs more attention. While these customers have higher monetary scores in comparison with the other segments, they have less purchase frequency and recency scores. This segment includes customers leaving the company. The customer churn should be noticed by the managers.

However, on the classification process, it was found that the company's dataset of this small E-business lacks some required data such as customer profile data. Concerning the data, the company was not well-organized due to the lack of long-term business strategy and required knowledge and experience. Therefore, Fuzzy Delphi approach has been employed to find the most important customer features we could use as predictors in classification models.

4. Fuzzy Delphi Application

As pointed out, it is possible to have a deficient database not allowing progress to goes in the classification process. This study proposes to use the fuzzy Delphi method to encounter the problem and gather the required data. Fuzzy Delphi method was derived from the traditional Delphi technique and fuzzy set theory (see Hsu et al. [9]). The traditional Delphi method is a structured communication technique, originally developed as a systematic, interactive forecasting method in which several rounds of anonymous written questionnaire surveys are conducted to ask for experts' opinion (Harold [8]). Noorderhaben [20] indicated that applying the fuzzy Delphi method to group decisions can appropriate the fuzziness of common understanding of expert opinions. Several researches have applied triangular fuzzy number, trapezoidal fuzzy number and Gaussian fuzzy number for fuzzy membership functions. Here, we use triangular membership function denoted by (m, α, β), where the point m, with membership grade of 1, is called the mean value and α, β are the left hand and right hand spreads of m, respectively.

To utilize the fuzzy Delphi method, the following steps were taken:

1- The relevant literature was studied to compile a list of variables and eventually eight variables as age, gender, education, income, purpose, lifestyle, occupation and location were derived.
2- Ten experts from different departments of company were invited and a questionnaire survey in the form of Table 2 was carried out. They were asked to determine the importance of each variable in identifying the customers for the classification purpose as very high, high, moderate, low, and very low. Table 2 shows the results. For instance, five experts evaluate the education as very high, four people evaluate it as high and one person evaluates it as moderate.
3- To show the experts' opinions as triangular membership functions, the linguistic variable scale (Fig. 2) and its equivalent triangular numbers (Table 3) were used. If the expert's opinion i for variable A is shown by

$$A_i = (m_i, \alpha_i, \beta_i). \quad (7)$$

Then, the fuzzy average of s experts' opinions for that variable is obtained by

Table 2. Experts' opinions about importance of variables in classification of customers

Variable		Agreement				
		Very High	High	Moderate	Low	Very Low
1	Location	0	0	2	5	3
2	Life Style	0	0	2	4	4
3	Education	5	4	1	0	0
4	Income	2	4	2	1	1
5	Purpose	5	2	2	1	0
6	Age	1	2	1	3	3
7	Gender	0	2	3	3	2
8	Occupation	0	1	2	4	3

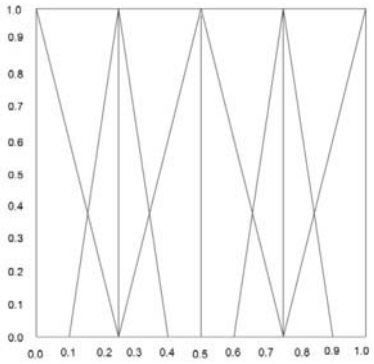

Figure 2. Linguistic variable scale

Table 3. Triangular fuzzy numbers and their equivalent linguistic variables (Mirsepasi et al., 2010).

Triangular Fuzzy Numbers and their Equivalent Linguistic Variables	
Linguistic Variable	Triangular number *(m, α, β)*
Very High	(1.00 ,0.25, 0.00)
High	(0.75, 0.15, 0.15)
Moderate	(0.50, 0.25, 0.25)
Low	(0.25, 0.15, 0.15)
Very Low	(0.00, 0.00, 0.25)

$$A_{ave} = (\frac{1}{s}\sum_{i=1}^{s} m_i , \frac{1}{s}\sum_{i=1}^{s} \alpha_i , \frac{1}{s}\sum_{i=1}^{s} \beta_i). \quad (8)$$

4- In order to rank the variables, the average of experts' opinions was calculated by the Minkowski formula given by

$$\chi = m + \frac{\beta - \alpha}{4}. \quad (9)$$

The fuzzy results are shown in Table 4. According to Table 4, the most important variables, education, purpose, income, gender and age, are used as predictors.

By interviewing customers of each segment using the phone, the required information about their age, gender, educational level, annual income and main purpose of internet usage was acquired. A total of 1808 customers were interviewed. These variables were quantified by encoding them as shown in Table 5. Considering these data along with age and income, there will be 13 variables as input for the classification routine. Table 6 indicates a summary of three distinct segments. In the following section, the classification process is made using this information and based on the aforementioned models.

Table 4. Variable selection by fuzzy Delphi

Features		Triangular Fuzzy Average (m, α, β)			De-fuzzyfied Average
1	Age	0.375	0.150	0.200	0.387
2	Gender	0.625	0.200	0.150	0.612
3	Education	0.850	0.210	0.085	0.819
4	Income	0.625	0.175	0.150	0.619
5	Purpose	0.775	0.220	0.095	0.744
6	Location	0.175	0.105	0.180	0.194
7	Life Style	0.100	0.06	0.200	0.135
8	Occupation	0.275	0.125	0.125	0.275

Table 5. Different levels of categorical variables

Variable	Code
Gender	Female=0, Mail=1
Education	Below Diploma=0, Diploma=1, University Diploma=2, B.S.=3, M.S.=4, Ph.D.=5
Purpose	Research=1, Business=2, Social Networking=3

Table 6. Statistics summary of three distinct segments

	Cluster 1	Cluster 2	Cluster 3
Gender	Female: 151 Male: 449	Female: 142 Male: 462	Female: 144 Male: 460
Education	Below Diploma:12 Diploma: 94 University Diploma: 76 B.S: 344 M.S.: 66 PhD: 9	Below Diploma: 3 Diploma: 73 University Diploma: 54 B.S: 345 M.S.: 116 PhD: 13	Below Diploma: 3 Diploma: 88 University Diploma: 87 B.S: 351 M.S.: 69 Ph.D: 6
Purpose	Research: 187 Business: 87 Social Networks and Entertainment: 327	Research: 256 Business: 177 Social Networks and Entertainment: 171	Research: 94 Business: 228 Social Networks and Entertainment: 282
Age	Mean:29.11 Median: 31 Min: 16 Max: 49	Mean:31.00 Median: 30 Min: 17 Max: 52	Mean: 32.91 Median: 31 Min: 17 Max: 55
Income	Mean:16.71 Median: 3 Min: 7 Max: 60	Mean: 22.98 Median: 20 Min: 9 Max: 85	Mean:16.71 Median: 3 Min: 7 Max: 72

5. Performance Evaluation

The performance of classification models can be measured by several criteria such as accuracy, hit rate, confusion matrix, etc.

Accuracy refers to the percentage of correct predictions made by the classification model when compared with the actual classifications in the validation data. A hit rate is the ratio of the number of correctly classified targets to the number of classified targets. A confusion matrix displays the number of correct and incorrect predictions made by the classification model compared with the actual classifications in the validation data. The matrix is n by n, where *n* is the number of classes. In order to use the different criteria, typically hold-out method is used in which the dataset is partitioned into two portions, training data and evaluation data. In other words, a percentage of the records are used to build the classification model and the remaining records were used to validate the model. In our study, for all the methods of classification, data records related to 1808 customers was divided using holdout sampling method into training and validation data sets with the ratio of 65 and 35 percent, respectively. Then, the accuracies of classification models were obtained in the Clementine and Matlab software environments and compared with each other.

5.1. The Results of LDF

As mentioned earlier, in order to use the method some criteria should be met such as normality of the sample distribution. The Shapiro Wilks test showed that except for age, the remaining variables are not normal. However, to get an insight of the Fisher's discriminant analysis, we ignore these assumptions. The aim of linear discriminant is to select the most significant variables and to determine the Fisher's linear discriminant function. Table 7 illustrates the most significant variables using the Lambda of Wilks (LW). LW is used in discriminant analysis such that the smaller the lambda for an independent variable, the more that variable contributes to the discriminant function. LW varies from 0 to 1, with 0 meaning that group means differ, and 1 meaning that all group means are the same (Dunteman [4]).

Table 8 shows that the most significant variables in differentiating the three customer segments in terms of Wilks' Lambda are income and purpose. To find the Fisher's linear discriminant functions as shown by (2), the coefficients of these two variables and the constant term should be determined. Table 5 shows the results. Therefore, we have three discriminant functions each of which discriminates one customer segment from the others as follows, according to equation 2:

$$F_1 = -5.506 + 0.074 \times \text{income} + 3.374 \times \text{purpose}$$
$$F_2 = -4.803 + 0.091 \times \text{income} + 2.862 \times \text{purpose}$$
$$F_3 = -7.319 + 0.132 \times \text{income} + 3.531 \times \text{purpose}$$

Table 7. The discriminant power of variables

Variable	Wilks' Lambda
Age	0.968
Sex	0.999
Education	0.979
Purpose	0.951
Income	0.866

Table 8. The coefficients of LDF functions

	Segments		
	1	2	3
Purpose	3.374	2.862	3.531
Income	0.074	0.091	.132
(Constant)	-5.506	-4.803	-7.319

Table 9. Correct classifications of LDF method

Classification Results					
	Class	Predicted Group Membership			Total
		1	2	3	
Count	1	206	145	44	395
	2	99	199	94	392
	3	113	89	172	374
%	1	52.2	36.7	11.1	100.0
	2	25.3	50.8	24.0	100.0
	3	30.2	23.8	46.0	100.0

To show the validity of the equations, the correct classifications for the validation data set are calculated. Table 9 exhibits the numbers and percentages of records classified correctly. True classifications into the first, second and third segments are respectively 52.2, 50.8, and 46 percent, yielding 49.7 percent on the average. The results show that Fisher's discriminant analysis performance is not adequate concerning correct classification criterion. It is because the model pre-assumptions are not verified

5.2. The Results of Logistic Regression (LR)

The logistic regression is a method not requiring any assumption. Therefore, it can be used in situations that the assumptions of linear discriminant procedure are not verified. To evaluate the LR model, discriminant functions are derived by training data and new observations are classified by the functions. If the segment one is considered as a reference group according to (1), then logistic regression finds two discriminant equations for segments two and three as follows:

F_3 = -4.096 + 0.4978 × [Sex=0] - 0.5183 × [Education=0] + 1.695 × [Education=1] + 1.972 × [Education=2] + 1.753 × [Education=3] + 1.54 × [Education=4] - 20.76 × [Purpose=0] - 0.2584 × [Purpose=1] - 0.1181 × [Purpose=2] + 0.1047 × Income

F_2= 0.0879 × [Sex=0] - 2.836 × [Education=0] - 0.4651 × [Education=1] - 0.9607 × [Education=2] - 0.8236 × [Education=3] - 0.981 × [Education=4] - 20.26 × [Purpose=0] + 0.9714 × [Purpose=1] + 0.7218 × [Purpose=2] + 0.0562 × Income - 0.8118

Classification of new observations and calculation of correct classifications can be done based on these equations, as shown in Table 10. The overall percentage of correct classification is 53.1%. Although this is a better value than that of Fisher's linear discriminant, it is still low.

5.3. The Results of Artificial Neural Network (ANN)

Different ANN methods use distinctive architectures or topologies including several hidden layers and various neurons in each hidden layer. To show the performance of artificial neural network, the aforementioned topologies considering different structures in employing training data are used. Table 11 indicates the performances of these methods along with their structures.

As shown, exhaustive prune method with the most complicated structure yields a better performance than that of others methods. Table 12 indicates the performances of ANN for both training and validation dataset. The average correct classifications of different structures are 58% and 60% for training and validation datasets, respectively.

Table 10. Correct classification of LR method

Classification				
Observed	Predicted			
	1	2	3	Percent age of Correct
1	231	121	43	58.5%
2	102	196	94	50.0%
3	113	72	189	50.5%
Overall Percentage	38.4%	33.5%	28.1%	53.1%

Table 11. Different ANN structured methods and their correct classifications percentage

Function	Input Layer Neurons	First Hidden Layer Neurons	Second Hidden Layer Neurons	Third Hidden Layer Neurons	Output Layer Neurons	Correct Classification (%)
RBFN	13	20	-	-	3	46.8
Dynamic	13	8	6	-	3	52.6
Exhaustive Prune	13	26	16	-	3	70.6
Multiple	13	12	12	11	3	64.4

Table 12. ANN's correct classifications for training and validation datasets

Method	Training		Validation	
RBFN	526	45%	308	48%
Exhaustive Prune	848	73%	481	74%
Dynamic	618	53%	354	55%
Multiple	725	62%	409	63%
Average	679	58%	388	60%

Table 13. SVM's correct classifications for training and validation datasets

Function	Training		Validation	
	Count	Percent	Count	Percent
RBF	675	58%	363	56%
Polynomial	841	72%	413	64%
Sigmoid	442	38%	239	37%
Linear	595	51%	353	54%
Average	638	55%	342	53%

5.4. Support Vector Machine (SVM)

SVM uses mathematical kernel functions to map the data into a high-dimensional feature space in order to categorize the dataset even when the data are not linearly separable. RBF, polynomial, sigmoid and linear kernel functions are used to classify the datasets and evaluate correct classifications for training and validation data. The results are shown in Table 13. Although polynomial function yields a better performance, however the average correct classifications are 55% and 53% for training and validation datasets respectively. In addition, the correct classifications of polynomial function for training and validation datasets are 72% and 64%, respectively. This shows a large difference and it is an indication of model instability.

5.5. Multi-group Discriminant Linear Programming

Based on the multi-group linear programming approach described in Section 2.5, and utilizing the formulations (3)-(6), the performance of the MLP model in classification is evaluated. For this purpose, a Matlab program has been prepared. The same hold-out sampling method with 65-35 percent for training and validation processes used.

5.5.1 Training Process

The first step is training the model and obtaining the weights of variables or predictors for each pair of segments according to formulation (4) and the introduced concepts in Section 2.5. The predictors in this study are the five demographic variables derived by a fuzzy Delphi method described in Section 3.1. The results are shown in Table 14, in which the W_i are the weight of age, gender, education, purpose and income, respectively. The acquired weights indicate the effect of each variable in classifying the customers between two segments. For instance, in classifying customer in segment 1 and 2, the weight of purpose (W_4) is the most important one, while the effect of age is not influential. This is true with classifying customers between segments 1 and 3 and segments 2 and 3 as well. This means that the purpose of internet usage, clarified as "doing research", "doing business" and "social networking and entertaining" strongly affect customer classification. When the results presented to the ISP company experts, were found to be consistent with their post experience. Now, the cut-off scores indicating the separating boundary between each pair of groups are obtained by Model 6. Table 15 shows the results.

Table 14. The weights of customer features in paired groups

Segments 1 and 2				
W_1	W_2	W_3	W_4	W_5
0.025721	0.093712	-0.11725	0.478386	-0.13682
Segments 1 and 3				
W_1	W_2	W_3	W_4	W_5
0.004049	0.086293	-0.01683	-0.12602	-0.062
Segments 2 and 3				
W_1	W_2	W_3	W_4	W_5
0.016368	0.012801	0.262658	-0.66177	-0.06957

Table 15. Cut off scores

	1	2	3
1		-0.9893	-1.32734
2			-1.89657
3			

Table 16. Predicted and current customer classes

Clusters 1 and 2				**Clusters 1 and 3**				**Clusters 2 and 3**			
ID	Score	Predicted Class	Current Class	ID	Score	Predicted Class	Current Class	ID	Score	Predicted Class	Current Class
3	-0.231	1	1	3	-0.888	1	1	601	-1.072	2	2
6	-2.552	2	1	6	-2.066	3	1	602	-2.036	3	2
7	-0.935	1	1	7	-0.754	1	1	604	-1.872	2	2
9	-0.342	1	1	9	-0.542	1	1	606	-0.491	2	2
12	-0.780	1	1	12	-1.219	1	1	607	-1.942	3	2
13	0.008	1	1	13	-1.213	1	1	610	-1.402	2	2
14	-1.029	2	1	14	-0.840	1	1	611	-3.916	3	2
15	-0.672	1	1	15	-1.517	3	1	613	-2.052	3	2
17	-0.989	1	1	17	-0.845	1	1	614	-2.477	3	2

As mentioned in Section 2.5, the deviations of individual objects (d_i) are minimized by the MDLP formulation. Therefore, having the weights of variables and cut-off scores and using (5), the classification scores of the training dataset are obtainable. Then, the corresponding class for each customer is predicted. Table 16 exhibits a small portion of the results. As shown, except for customers 6 and 14, which belong to segment 1 but have been assigned to wrong groups, the other customers are assigned correctly. The percentages of correct classifications for paired groups are shown in Table 17. It shows that correct classification between cluster 1 and 3 is the highest and the average is 66%. This performance is near the best results obtained for unique structures of artificial neural networks.

Table 17. The correct classification percentages of MDLP method for training data

Cluster Comparison	Correct Classification
1 & 2	65%
1 & 3	69%
2 & 3	63%
Average	66%

Table 18. The correct classification of MDLP method for testing data

Cluster Comparison	Correct Classification
1 & 2	68%
1 & 3	73%
2 & 3	65%
Average	69%

5.5.2. Validation

Using cut-off and weight scores obtained in the previous stage, the classification scores of validation dataset are obtained. Table 18 shows the result.

Table 18 indicates that the correct classification for validation data is slightly better than correct classification for training data. However, the average correct classification is 69% and it is very near to the average correct classification for training data, showing the stability of the model.

5.6. Interval Multi-group Discriminant Linear Programming

The process used for MDLP is now implemented using models (9) and (10). A Matlab code was written enabling user to input uncertainties for the variables. Among five variables, namely gender, age, education, annual income and purpose, it is believed that uncertainty may impress more annual income than other variables. Question about people's income being sensitive, a 20 percent spread is considered for the expressed income values, i.e., the lower and upper bounds are respectively 80 and 120 percent of the exact value.

The obtained IMDLP results reveal two main differences with respect to the MDLP results. First according to Table 19, the objective function value of IMDLP is almost halved, in comparison to MDLP. Second, the classification performance of IMDLP is relatively better than that of MDLP for training and testing datasets, as exhibited in tables 20 and 21

Table 19. MDLP and IMDLP objective function values

The value of MDLP objective function	The value of IMDLP objective function
473.90	266.49

Table 20. Comparison MDLP and IMDLP results for training dataset

Training Dataset				
	Classes 1,2	Classes 1,3	Classes 2,3	Average
MDLP	65.50%	69.48%	63.39%	66.12%
IMDLP	67.22%	71.19%	66.11%	68.17%

Table 21. Comparison MDLP and IMDLP results for testing dataset

Testing Dataset				
	Classes 1,2	Classes 1,3	Classes 2,3	Average
MDLP	68.17%	73.02%	65.39%	68.86%
IMDLP	69.22%	74.19%	66.11%	69.84%

Table 22. Comparison of performance of different classification methods

Model	Training	Testing
ANN	58%	60%
LDF	50%	53%
LR	53%	55%
SVM	55%	53%
MDLP	66%	69%
IMDLP	68%	70%

5.7. Comparison of the Performance of the Models

This section compares the performance of the aforementioned classification models. Table 22 includes the average correct classifications of the models for training and validation datasets. As shown, the average correct classification of MDLP and specially of IMDLP is considerably higher than the ones for other models. However, a major concern for application of multi-group linear programming is its of relatively high running time requirement. Using an Intel Pentium Dual CPU and E2200 @ 2.2 GHz PC computer with 2G RAM, the method is slightly slower than the other models and takes 3 minutes and 40 seconds to run the LP model for 1808 record datasets. This amount of time is comparable with the time required by other models.

6. Conclusion

Statistical and data mining classification methods have been used for years. Although different mathematical programming models for classification have been introduced in the literature, but they have been ignored due to their considerable computing costs. However, the advent of powerful information systems has resumed the application of mathematical programming due to its non-parametric nature of not requiring on any assumption. Our study used the well-known statistical and data mining methods vis-à-vis multi-group discriminant linear programming in a real setting of an ISP company. The real life data often is contaminated, because it usually includes noisy, incomplete and redundant information. Specially, the problem of lack of required data is an important issue. To overcome this problem, we used fuzzy Delphi to extract the necessary data for classification purposes. In addition, to absorb the uncertainties imposed on the collected data, we developed the

interval MDLP. The results showed that the average performances of MDLP and IDMLP, in particular, are considerably preferred, at least for small and medium-sized datasets.

Acknowledgments

The authors would like to thank the Irangate ISP company for providing the required data.

References

[1] Birant, D. (2011), Data Mining Using RFM Analysis, In K. Funatsu and Hasegawa, K. (Eds.), Knowledge Oriented Applications in Data Mining (pp. 91-108), Rijeka, Croatia: InTech.

[2] Blattberg, R.C., Kim, B., and Neslin, S.A. (2008), Database Marketing: Analyzing and Managing Customers, New York: Springer.

[3] Celebi, D. and Bayraktar, D. (2008), An integrated neural network and data envelopment analysis for supplier evaluation under incomplete information, *Expert Systems with Applications*, 35, 1698-1710.

[4] Dunteman, G. (1984), Introduction to Multivariate Analysis, Thousand Oaks, CA, Sage Publications.

[5] Dyche, J. and Dych, J. (2001), The CRM handbook: A Business Guide to Customer Relationship Management, Reading, MA: Addison-Wesley.

[6] Flach, P. (2001), On the state of the art in machine learning: A personal review, *Artificial Intelligence*, 131(1-2), 199-222.

[7] Freed, N. and Glover, F. (1981), Simple but powerful goal programming models for discriminant problems, *European Journal of Operational Research,* 7, 44-66.

[8] Harold A., M.T. (1975), The Delphi Method: Techniques and Applications, Reading, Addison-Wesley.

[9] Hsu, Y. L., Lee, C.H., and Kreng, V.B. (2010), The application of fuzzy Delphi Method and fuzzy AHP in lubricant regenerative technology selection, *Expert Systems with Applications*, 37, 419-425.

[10] Johnson, R. and Wichern, D. (1988), Applied Multivariate Statistical Approach, Englwood Cliffs, NJ, Prentice-Hall.

[11] Kiang, M.Y., Hu, M.Y. and Fisher, D.M. (2006), An extended self-organizing map network for market segmentation-telecommunication example, *Decision Support Systems*, 42, 36-47.

[12] Kotsiantis, S. and Pintelas, P. (2004), Recent advances in clustering: a brief survey, *WSEAS Transactions on Information Science and Applications*, 1, 73-81.

[13] Lam, K., Choo, E. and Moy, J. (1996), Improved linear programming formulations for the multi-Group discriminant problem, *Journal of the Operational Research Society*, 47(12), 1526-1529.

[14] MacQueen, J. (1967), Some Methods for Classification and Analysis of Multivariate Observations, Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability, 1, University of California Press, Berkeley, pp. 281-297.

[15] Mangasarian, O. (1965), Linear and nonlinear separation of patterns by linear programming, *Journal of Operations Research*, 13, 444-452.

[16] McCarty, J. and Hastak, M. (2007), Segmentation approaches in data-mining: A comparison of RFM, CHAID, and logistic regression, *Journal of Business Research*, 60, 656-662.

[17] Meyers, L., Gamst, G. and Guarino, A. (2006), Applied Multivariate Research: Design and Interpretation, Thousand Oaks, CA: Sage Publications, Inc.

[18] Michie, D., and Spiegelhalter, D. (1994), Machine Learning, Neural and Statistical Classification, Taylor.

[19] Morrison, D. (1969), On the interpretation of discriminant analysis, *Journal of Marketing Research*, 6, 156-163.

[20] Noorderhaben, N. (1995), Strategic Decision Making, UK, Addison-Wesley.

[21] Pai, D.R., Lawrence, K.D., Klimberg, R.K. and Lawrence, S.M. (2012), Experimental comparison of parametric, non-parametric, and hybrid multigroup classification, *Expert Systems with Applications*, 39, 8593-8603.

[22] Shmueli, G., Patel, N. and Bruce., P. (2006), Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner, NJ, John Wiley and Sons, Inc.

[23] Sun, M. (2010), Linear Programming approaches for multiple-Class discriminant and classification analysis, *International Journal of Strategic Decision Sciences*, 1(1), 57-80.

[24] Vapnik, V. (1995), The Nature of Statistical Learning Theory, NY, Springer.

[25] Witten, I. and Frank, E. (2005), Data Mining, Practical Machine Learning Tools and Technique, Oxford, UK, Elsevier.

[26] Youssef, S. and Rebai, A. (2007), Comparison between statistical approaches and linear programming for resolving classification problem, *International Mathematical Forum*, 63, 3125-3141.

[27] Ziniel, S. (2010), Avoiding Bias in the Research Interview, Childrens' Hosptial, Boston, Harvard Medical School.

[28] http://pic.dhe.ibm.com/infocenter/spssmodl/v15r0m0/index.jsp?topic=%2Fcom.ibm.spss.modeler.help%2Fneuralnet_modeltab.htm.