Proceedings of the International Conference

on

MODELING AND SIMULATION

General Application & Embedded Systems Engineering

The Association for Modeling and Simulation in Enterprises (AMSE) and Selçuk University

with the Selçuk University Research Center of Applied Mathematics

August 28 – 30 , 2006

Konya, TURKEY

Editor

Prof. Dr. Novruz ALLAHVERDI

Assistant Editors

Assoc. Prof. Dr. Aşır GENÇ

Assis. Prof. Dr. Galip OTURANÇ

INTERNATIONAL CONFERENCE ON MODELLING AND SIMULATION

General Application and Embedded Systems Engineering (AMSE 2006)

August 28 – 30, 2006, Konya, TURKEY

HONORARY COMMITTEE

Prof. Süleyman OKUDAN, Turkey - Rector of Selcuk University

Prof. Jaime Gil ALUJA, Spain - President of AMSE

INVITED SPEAKERS

Prof. Christian BERGER-VACHON, France

Prof. Victor KRASNOPROSHIN, Belarus

Prof. Burhan TÜRKŞEN, Canada; Turkey

ORGANIZING COMMITTEE

Prof.Dr. Novruz Allahverdi

Prof.Dr. Saadettin Herdem

Dr. Galip Oturanç

Dr. Şirzat Kahramanlı

Dr. Aşır Genç

Dr. Mustafa Acaroğlu

Dr. Ulfet Atav

Dr. Oğuz Doğan

Dr. Cengiz Çınar

Dr. Aydın Kurnaz

Dr. Mustafa Semiz

Dr. Süleyman Tosun

Dr. Orhan Engin

Dr. Ramazan Türkmen

Dr. Ozan Özkan

Dr. Hakan Işık

Dr. Yüksel Özbay

Dr. Mehmet Çunkaş

Dr. Ali Kahraman

Dr. Cemil Sungur

Kemal Tütüncü

Alpaslan A. Altun

Yıldıray Keskin

Tarık Yılmaz

M. Nevzat Örnek

SCIENTIFIC COMMITTEE

Prof. José Casajuana Gibert, Spain

Prof. Nashaat El-Khameesy, Egypt

Prof. Prabhat Kumar Mahanti, India

Prof. Christian Berger-Vachon, France

Prof. Joan Carles Ferrer, Spain

Prof. Bernard Ballaz, France

Prof. KeDi Huang, China

Prof. Anna-Maria Gil Lafuente, Spain

Prof. Ramiz Mirsalimov, Azerbaijan

Prof. Victor Krasnoproshin, Belarus

Prof. Hristo Dalkalatchev, Bulgaria

Prof. Basel Al-Eideh, Kuwait

Prof. Dragutin Debeljkovic, Serbia

Prof. Rafael Espin, Cuba

Prof. Houcine Chafouk, France

Prof. D.P. Kothari, India

Prof. Francesco Carlo Morabito, Italy

Prof. Vladimir S. Neronov, Khazakstan

Prof. Said Allaki, Morocco

Prof. Andrzej Horodecki, Poland

Prof. Vasyl Tchaban, Ukraine

Prof. Jaime Tinto Arandes, Venezuela

Prof. İsmihan Bayramoğlu, Turkey

Prof. Ahmet Arslan, Turkey

Prof. Durmuş Bozkurt, Turkey

Prof. Seyhan Uygur Onbaşıoğlu, Turkey

Prof. Ahmet M.Gökçen, Turkey

Prof. Hamza Erol, Turkey

Prof. Fikri Akdeniz, Turkey

Prof. Ömer L.Gebizlioğlu, Turkey

Prof. Serdar Kurt, Turkey

Prof. Ali Ünüvar, Turkey

Prof. Burhan Türkşen, Turkey

Prof. Fikret Ercal, USA

Prof. Gürbüz Atagündüz, Turkey

Prof. Ömer Akın, Turkey

Prof. Necdet Özbalta, Turkey

Prof. Mustafa Güneş, Saudi Arabia

Prof. Mehmet Bayrak, Turkey

Prof. Kayhan Erciyeş, Turkey

Prof. İnan Güler, Turkey

Prof. Mehmet Pakdemirli, Turkey

Prof. Ö. Ercan Ataer, Turkey

Prof. Bekir Karlık, Turkey

Prof. Kemal Altınışık, Turkey

B209	Acoustic echo cancellation with adaptive filtering	Y. OZBAY, A. R. KAVSAOGLU	649-653
B125	Fuzzy clustering based c-linear regression modeling	M. EMINOV, N. GULER	655-660
B164	Constitution of stability domain for handing asymptotic stable scaler equations	O. A. GUMUŞ, H. KOSE	661-664
B022	A new approach to distribution of frequency histogram of discontinuies spacing in rock mass structure with fuzzy logic method	A. TURANBOY, Y. E. GOKTEPE ,Y. UZUN	665-670
A025	Multifractal, genetic algorithms and neural network in cervical cell images recognition	L. HAMAMI, N. Z. LASSOUAOUI,N NOUALI, H. SAADI	671-676
B005	ICVS: Intelligent caching of video segments over IEEE802.11 wireless lan	M. WAHEEB, I. AYBAY	677-681
B067	Voip simulation on ttnet : The case of TURKIYE	M. GUNES,O. AYDOGAN	683-690
B101	Modelling and simulation of wireless sensor and mobile ad hoc networks	K. ERCIYES, O. DAGDEVIREN	691-695
B119	Vectorization of raster images while keeping their original width	M. GUNDUZ, F. YILDIZ, H. KARABORK	697-700
B010	The advantages and disadvantages of expert systems applications in managerial processes: The case of tourism industry	A. OGUT, T. AKGEMCI, T. GOK	701-706
B011	The development of corporate performance management (CPM) with data mining modeling in the enterprises	A. OGUT, R. PARLAKKAYA, A. KOCABACAK	707-711
A124	Oil price modeling considering us$/euro exchange rate using anfis	A. GHAFFARI, M. NIKKHAH BAHRAMI, S.M GHAFFARI	713-718
A106	Acquisition of financial products that adapt to different environments	J.M. MERIGO, A.M. GIL LAFUENTE	719-723
A114	Using the owg operators in the selection of financial products	J.M. MERIGO, A.M. GIL LAFUENTE	725-728
B162	Evaluation of risky investment projects through simulation and response surface methodology	O. ARMANERI, O. YALCINKAYA	729-734
B021	Synthetic generation of rainfall data using Thomas-Fiering model	A. I. MARTI, U. SENDIL	735-740
B129	A comparison of nonhierarchical clustering methods	M. ERISOGLU, A PEKGOR, A. KARAKOCA, A. GENC	741-744
B157	Modelling of persuasion among decision makers for AHP (analytic hierarchy process) based decision making process	O. OZVERI	745-749
A042	A general framework for fuzzy multiple attribute decision- making problem	F.F.FARAHAT	751-756
B095	Economic effects of current account deficits: Simulation based approximations	I. AKGUL, M. K. BESER	757-763
A043	A new efficient rapid algorithm for performance evaluation for large scale HS TDS networks for BISDN	F.F. FARAHAT	765-769
B027	Ordinary and penalized power-divergence test statistics in two-way contingency tables	A. ALIN, S. KURT	771-776
B049	Seasonal demand forecasting in a multi-product setting: a case study	D. T. ELIIYI, A. F. CELIK, E. OZLEM, H. S. YUKSEL	777-782
B147	Fuzzy multilayer perceptron for fuzzy regression	N. Y. PEHLIVAN, A. APAYDIN	783-787
B148	On the comparison of median smoothing and radial basis function networks	N. Y. PEHLIVAN, A. APAYDIN	789-793
B168	Identifier of initial conditions for adjoint variables in minimum time optimal control problem	R.Y. SAMEDOV, Q.A. RUSTAMOV, M.B. NAMAZOV, V.Q. FARHADOV	795-797
B223	Agricultural impact analysis towards Turkey's accession to European Union: An evaluation study	A.MERIC, A.. UZUNPINAR, M. ASLAN	799-803
B025	On some measures of nonlinearity in regression analysis	A. EVREN , E. OZTURK	805-810
B026	On the comparison of some mutivariate classification techniques by using heterogeneity measures such as entropy and Gini index	A. EVREN, D YILDIZ	811-816
B038	Bayes estimation with non-informative priors for stress-strength reliability under Burr XII distribution	I. KINACI, C. KUS, B. SARACOGLU	817-820

MS2006
Proceedings of the International Conference on Modeling and Simulation 2006
28 – 30 August 2006, Konya, TURKEY
Paper No: B026

ON THE COMPARISON OF SOME MUTIVARIATE CLASSIFICATION TECHNIQUES BY USING HETEROGENEITY MEASURES SUCH AS ENTROPY AND GINI INDEX

Atıf Evren Doğan Yıldız
YTU, Faculty of Art and Science, Departmant of Statistics,
Davutpaşa-Esenler İSTANBUL
aevren@hotmail.com doganstar@gmail.com

ABSTRACT

Entropy which plays a crucial role in information theory, is used to determine the degree of variability of a probability (or a frequency) distribution. The larger the variability (or heterogeneity) within the probability distribution, the larger will be entropy. Thus in case of large entropy, the mean of the probability distribution will be a poor estimate of the central tendency.
Especially, when the random variable is qualitative, it is impossible to calculate the mean of a distribution. Hence the variance as a measure of variability can not be obtained. Therefore it is necessary to use mode as a measure of central tendency and to use entropy as a measure of dispersion instead.
While the lower limit of entropy is zero, the upper limit depends on the number of groups of a distribution. The more the number of groups, the bigger the entropy. The tendency to increase indefinitely as the number of groups increase causes some criticisms about entropy
As an alternative to entropy, a modified version of Gini index might be computed. In literature, Gini index is a measure to determine the degree of inequality of an income distribution. Apart from the original context of Gini index, a modified version might be used to compute the degree of variability of a probability distribution. The vertical scale of the diagram for Lorentz curve might be used to represent the cumulative probabilities. Again the line for "absolute equality" might be drawn. Thus similar reasoning can be followed by comparing the area formed under the line for absolute inequality and the area formed by interpolating the cumulative probability points on the diagram. Here again, the greater the variability of a probability distribution, the greater will be Gini index.
The classification techniques such as cluster analysis and discriminant analysis group data according to certain similarity measures. The success of a such grouping technique depends on how homogeneous will be each group after classification.
In the empirical part of this study, the customers of a certain banking institution will be classified in terms of their multivariate credibility data .(according to credit scores of customers) The customers grouped under one extreme will be the "good" ones, while the ones grouped at the other extreme will be specified as the "bad" ones. After such classifications are realized under different grouping techniques, the success of each grouping technique will be evaluated by using homogeneity (or heterogeneity) measures such as entropy, Gini index or similar measures.

Keywords: Entropy, Gini Index, Multivariate Classification Techniques.

1.1 ENTROPY

Let **X** be a random variable(whether discrete or continuous) , be an experiment, **S** the sample space generated by performing and **f(X)** the probability (or probability density) function of **X**. Shannon and Wieder proposed the notion of entropy as a measure of variability within a probability distribution.* Entropy of a distribution is defined as the negative of the expected value of **log [f(X)]**.

If **H(X)** denotes the entropy of a probability distribution then

H(X)=E[-log(f(X)] (1.1.1)

If **X** is discrete and takes the values $x_1, x_2, \cdots, x_n$ on **S**,

$$H(X) = -\sum_{i=1}^{n} f(x_i)\log(f(x_i)) \qquad (1.1.2)$$

and if **X** is continuous and sample space **S** is simply equivalent to the set of real numbers $\Re$,then

$$H(X) = -\int_{-\infty}^{\infty} f(x)\log(f(x))dx \qquad (1.1.3)$$

Here the base of the logarithm is immaterial. But for the practical purposes of communication engineering, the base is generally taken as 2.

In discrete probability distributions the term $\log(f(x_i))$ is generally evaluated as the measure of entropy (or uncertainty) related to the event $\{X = x_i\}$ in sample space **S**. If $\{X = x_i\}$ is a sure event then $P\{X = x_i\} = 1$ and $\log(f(x_i)) = 0$ so there is no uncertainty stuck on the event $\{X = x_i\}$ in **S**. On the other hand if $P\{X = x_i\} \cong 0$ then the term $\log(f(x_i))$ will tend to infinity in its absolute value which is of course corresponds to the case with biggest entropy. The more probable the event will take place in **S**, the lower the level of entropy. The less probable the event in **S**, the bigger the level of entropy.
Since the random variable takes the values $x_1, x_2, \cdots, x_n$ on **S**, a more "average" measure to consider the whole entropy within a probability distribution is needed. It might be based on the expected value of all $\log(f(x_i))$ values. Therefore the above definitions of entropy are aimed to measure average entropy.

* Reza, F.M.; "An Introduction to Information Theory", Dover Publications, pp77.

For the discrete distributions the minus sign in front of the summation sign guarantees that the entropy is a positive quantity whereas for the continuous distributions the situation is slightly different. The first difference is that in continuous distributions the probability density function **f(x)** might be bigger than 1. In such a case the entropy could be negative. Secondly f(x) is not the probability that **X** is equal to **x**. Rather **f(x)** is a measure of probability density at **X=x.** These differences must be taken into consideration while dealing with continuous probability distributions.

Suppose **X** is discrete and takes the values $x_1, x_2, \ldots, x_n$ on **S**. The greatest entropy occurs, when all elementary outcomes $\{X = x_i\}$ i=1,2,...,n in **S** is equiprobable. In this situation by the axioms of probability,

$$P(X = x_i) = \frac{1}{n} \qquad i=1,2,\ldots,n \qquad (1.1.4)$$

And the corresponding (maximum) entropy is attained at the level of

$$H(X) = -\sum_{i=1}^{n} \frac{1}{n} \log(\frac{1}{n}) = \log n \qquad (1.1.5)$$

On the other extreme, let us suppose that the event $\{X = x_k\}$ in **S** is sure event such that $P\{X = x_k\} = 1$ k=0, 1, 2,..., n and $P\{X = x_i\} = 0 \quad i \neq k$ then

H(X)=0

Thus for the discrete probability distributions the entropy is between 0 and **log n.**

1.2 Multivariate Distributions

Let **(X, Y)** is a bivariate random variable on **S**. Suppose that the joint probability distribution of **(X, Y)** is $f_{X,Y}(x,y)$ In this case if **H(X, Y)** is the joint entropy, and if **X** and **Y** are discrete,

$$H(X,Y) = -\sum_{j=1}^{m}\sum_{i=1}^{n} f_{X,Y}(x_i,y_j)\log\left(f_{X,Y}(x_i,y_j)\right) \qquad (1.2.1)$$

if X takes the values $x_1, x_2, \ldots, x_n$ and **Y** takes the values $y_1, y_2, \ldots, y_m$ on sample space **S**.

If **(X, Y)** is continuous, the joint entropy is defined as

$$H(X,Y) = -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\log\left(f_{X,Y}(x,y)\right)dxdy \qquad (1.2.2)$$

here the sample space **S** is equivalent to $\Re^2$.

For the n-variate cases generalizations are straightforward:

For the discrete cases;

$$H(X_1, X_2, \ldots, X_n) = -\sum_{X_1}\sum_{X_2}\ldots\sum_{X_n} \log(f(x_1,x_2,\ldots,x_n))f(x_1,x_2,\ldots,x_n) \qquad (1.2.3)$$

For the continuous cases;

$$H(X_1, X_2, \ldots, X_n) = -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\ldots\int_{-\infty}^{\infty} \log(f(x_1,x_2,\ldots,x_n))f(x_1,x_2,\ldots,x_n)dx_1dx_2\ldots dx_n \qquad (1.2.4)$$

1.3 Conditional Entropies

Conditional entropies are calculated by using conditional probability distributions. Consider the two conditional probability functions $f_{X/Y}(x/y)$ and $f_{Y/X}(y/x)$. One can calculate conditional entropies since

$$f_{X/Y}(x/y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}; \quad for \quad f_Y(y) \neq 0 \qquad (1.3.1)$$

$$f_{Y/X}(y/x) = \frac{f_{X,Y}(x,y)}{f_X(x)}; \quad for \quad f_X(x) \neq 0 \qquad (1.3.2)$$

Thus

$$H(X/Y) = -\sum_{i=1}^{n} f_{X/Y}(x_i/y)\log(f_{X/Y}(x_i/y)) \qquad (1.3.3)$$

and

$$H(Y/X) = -\sum_{j=1}^{m} f_{Y/X}(y_j/x)\log(f_{Y/X}(y_j/x)) \qquad (1.3.4)$$

Of course if the distributions are continuous, the summation operators in the above definitions must be replaced by integral operators. Now for all $\{Y = y_j\}$ events if the average conditional entropy of **X** is calculated and this quantity be denoted by $H(\bar{X}/Y = y_j)$ then

$$H(\bar{X}/Y = y_j) = \sum_{j=1}^{m} H(X/Y = y_j).f_Y(y_j) \qquad (1.3.5)$$

$$= -\sum_{j=1}^{m}\sum_{i=1}^{n} f_{X,Y}(x_i,y_j)\log\frac{f_{X,Y}(x_i,y_j)}{f_Y(y_j)} \qquad (1.3.6)$$

Similarly

$$H(\bar{Y}/X = x_i)$$

$$= -\sum_{j=1}^{m}\sum_{i=1}^{n} f_{X,Y}(x_i,y_j)\log\frac{f_{X,Y}(x_i,y_j)}{f_X(x_i)} \qquad (1.3.7)$$

Here H**(X)** shows the entropy in the marginal probability distribution of **X.**

H(X,Y) measures the entropy of the joint distribution whereas $H(X/Y = y_k)$ measures that of the conditional distribution of **X** given the event $\{Y = y_k\}$.

1.4 Entropy and Statistical Independence

It is important to note that when the variables **X** and **Y** are independent, the joint probability distribution is the product of two marginal probability distributions. It follows that in case of independence of two random variables;

H(X/Y) =H(X) and **H(Y/X) =H(Y)**

Both of which mean that the occurrence of one of the two independent events $\{X = x_i\}$ or $\{Y = y_k\}$ do not give any information about the entropy of the probability distribution of the other.

Besides it is possible to set up some relationships between different entropy types by probability rules.

For example it is easy to show that

H(X,Y)=H(Y)+H(X/Y) or **H(X,Y)=H(X)+H(Y/X)**

Here, it is interesting to note that the joint entropy might be viewed as the sum of a marginal type of entropy and a conditional type of entropy.

1.5 A Measure of Mutual Information

The average amount of information conveyed by the random variable **Y** about **X** was suggested by C. E. Shannon[*]. It is defined as

$I(X,Y)$

$$-\sum_{i=1}^{n}\sum_{j=1}^{m} P(X=x_i, Y=y_j)\log\frac{P(X=x_i, Y=y_j)}{P(X=x_i).P(Y=y_j)} \qquad (1.5.1)$$

Here the random variables are both discrete. For continuous cases the above definition may be adapted easily by replacing summation operators with integration operators. If **X** and **Y** are independent it is evident that $P(X=x_i, Y=y_j) = P(X=x_i).P(Y=y_j)$ for all x_i and y_j. If this is the situation, then I(X,Y)=0 which means that if the random variables are independent the average amount of information given by one random variable about the other is zero. It should be noted that $I(X,Y) \geq 0$

While the mutual information quantities may become negative for some pairs of events $\{X=x_i\}$ and $\{Y=y_j\}$ in **S**. By direct application of the defining equations it can be shown that

I(X,Y)=H(X)+H(Y)-H(X,Y) (1.5.2)

or

I(X,Y)=H(X)-H(X/Y) (1.5.3)

or

I(X,Y)=H(Y)-H(Y/X) (1.5.4)

Thus two concepts, namely, entropy and mutual information are related to each other.

Some applications of the entropy theory can be found in communication theory in electrical engineering. In statistical theory, as mentioned above, entropy calculations might be useful while analyzing the dependencies of random variables. When the variable(s) are qualitative neither the arithmetic mean can be used as a measure of central tendency nor can the variance be used as a measure of variability of a distribution. In such circumstances, the mode (as a measure of central tendency) and the entropy (as a measure of variation) can be helpful.

2 Gini Index

Gini coefficient is a measure of inequality proposed by Corrado Gini and is generally used to measure income inequality. Gini coefficient is between 0 and 1, where 0 corresponds with perfect equality (where everyone has the same income) and 1 corresponds with perfect inequality (where one person has all the income, and everyone else has zero income).The **Gini index** is the Gini coefficient expressed in percentage form, and is equal to the Gini coefficient multiplied by 100.

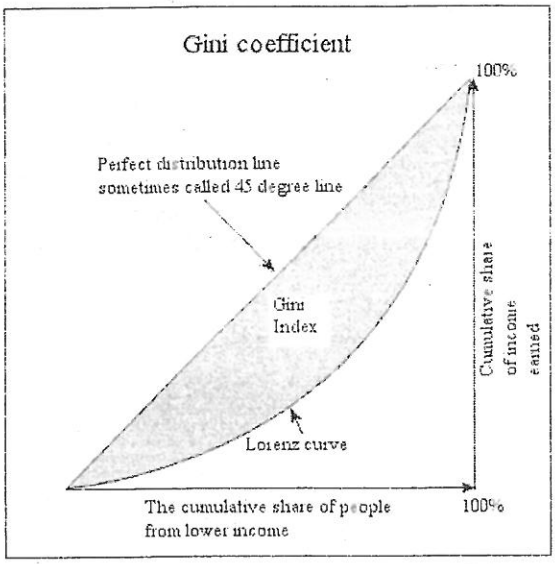


Figure 2.1: Graphical representation of the Gini coefficient[*]

The Gini coefficient is calculated as a ratio of the areas on the Lorenz curve diagram. If the area between the line of perfect equality and Lorenz curve is A, and the area underneath the Lorenz curve is B, then the Gini coefficient is A/(A+B). The Gini coefficient is often calculated with the more practical Brown Formula;

$$G = \left|1 - \sum_{k=1}^{n}(X_k - X_{k-1})(Y_k + Y_{k-1})\right| \qquad (2.1)$$

G:Gini coefficient

$\mathbf{X_k}$: cumulated proportion of the population variable,

for **k = 0,...,n**, with $\mathbf{X_0 = 0}$, $\mathbf{X_n = 1}$

$\mathbf{Y_k}$: cumulated proportion of the income variable,

for **k = 0,...,n**, with $\mathbf{Y_0 = 0}$, $\mathbf{Y_n = 1}$

It has been shown that the sample Gini coefficients defined above need to be multiplied by $n/(n-1)$ in order to become unbiased estimators for the *population* coefficients[*]

By similar reasoning an uncertainty within a probability distribution can be analyzed approximately. Consider a discrete random variable **X** which takes the values $x_1, x_2, \ldots, x_n$[**] with probabilities $f(x_1), f(x_2), \ldots, f(x_n)$ correspondingly.

Suppose the values $x_1', x_2', \ldots, x_n'$ are chosen among $x_1, x_2, \ldots, x_n$ such that x_1' is the **X** value whose probability is the smallest, x_2' is the **X** value whose probability is the second smallest;...; x_n' is the **X** value whose probability is the largest. Suppose also that after this ordering, the cumulative probabilities are

---

[*] Reza, F.M.: "An Introduction to Information Theory", Dover Publications

[*] The extensions to the cases in which there are more than two variables can be realized directly. For a formal proof of the equations above see Khinchin (Mathematical Foundations of the Information Theory: Dover Publications, New York 1957)

[*] http://mathworld.wolfram.com/GiniCoefficient.html

[**] If the data is grouped then $x_1, x_2, \ldots, x_n$ can be taken as the means (or the midpoints) of each group.

$F(x_1'), F(x_2'), \ldots F(x_n')$

If a graph displaying the relationship between X' and $F(X')$ is obtained a similar reasoning might be made. If all $\{X = x_i\}$ events take place with equal probability, the maximum uncertainty is met. If there are **n** groups (or if the random variable **X** takes n different values) with equal probability, the graph obtained is as follows:

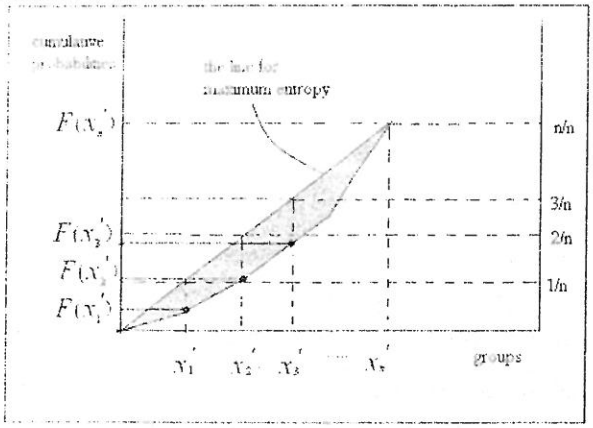


Figure 2.2: Diagram for the Calculation of the Degree of Gini type of Heterogeneity of a Distribution

In this diagram, the maximum area formed under the curve is the area of the triangle and is equal to $0{,}5x_n'$ this situation corresponds to the maximum disorder (entropy). If any one of the $\{X = x_i\}$ event in **S** takes the probability 1, then the area under such a curve is equal to 0. Thus for specific **X** values, the magnitude of minimum disorder (which is equal to 0) and the magnitude of maximum disorder (which is equal to $0{,}5x_n'$) can be determined by calculating the area under the curve obtained. This area may be approximated by the formula

$$Area = \sum_{i=1}^{n} \frac{[F(x_i') + F(x'_{i-1})]}{2} [x'_i - x'_{i-1}] \qquad (2.2)$$

where $F(x'_0) = 0$ and $x'_0 = 0$*

3 Multivariate Classification Techniques

An important task of all scientific studies is the classification of the phenomena being studied.** Classification models are used to categorize an object on the basis of a profile of characteristics. Let **z:** px1 denote a vector of observed attributes of an object or individual, and suppose it is known that **z** is an observation from exactly one of the mutually exclusive populations 1, 2, K. The problem is to discriminate among the populations; that is, a decision must be made as to which population **z** belongs.***

There is a vast literature on classification and discrimination. In order to classify **z** into one of the populations, Fisher (1936) suggested as a basis for classification decisions the use of a discriminant function linear in the components of **z**. Other bases for classification have included likelihood ratio tests, information theory and Bayesian techniques. Only two of the general techniques will be applied in this paper.

Discriminant analysis is used in situations where the clusters are known a priori. The aim of discriminant analysis is to classify an observation, or several observations, into these known groups. For instance, in credit scoring, a bank knows from past experience that there are good customers (who repay their loan without any problems) and bad customers (who showed di_culties in repaying their loan). When a new customer asks for a loan, the bank has to decide whether or not to give the loan. The past records of the bank provide two data sets: multivariate observations xi on the two categories of customers (including for example age, salary, marital status, the amount of the loan, etc.). The new customer is a new observation x with the same variables. The discrimination rule has to classify the customer into one of the two existing groups and the discriminant analysis should evaluate the risk of a possible bad decision.

Cluster analysis is a technique used for combining observations into groups or clusters such that each group is homogeneous and each group should be different from other groups with respect to the same characteristics.* As a special case, we apply the two step cluster analysis procedure. This technique is based on the assumption that the variables are independent; a joint multinomial -- normal distribution can be placed on categorical and continuous variables.

4. Application.

In the empirical part of this study, the customers of a certain banking institution will be classified in terms of their multivariate credibility data .(according to credit scores of customers) The customers grouped under one extreme will be the "good" ones, while the ones grouped at the other extreme will be specified as the "bad" ones. After such classifications are realized under different grouping techniques, the success of each grouping technique will be evaluated by using homogeneity (or heterogeneity) measures such as entropy, Gini index or similar measures. The sample size is **n=1422.**

Each customer has the following entries: Age, gender, customer type, salary, net monthly income, car ownership, other credit card ownership, and employment type. There are three quantitative variables used in this analysis: age, salary and net monthly income.

As a first classification method, discriminant analysis has been performed by using these three variables as discrimination factors .The following table gives the code numbers of each variable.

---

* This formula is obtained by summing the areas of triangle and trapezoids under the curve in the figure above.

** Everitt.B.S.:G.Dunn, Applied Multivariate Data Analysis, Arnold Publishers. Second Edition.2001. pp125

*** Press. S.J..Applied Multivariate Analysis Using Bayesian and Frequentist Methods of Inference. Dover Publications. Second Edition.. 2005.pp396.

**** Hardle.W.;Simar,L.: Applied Multivariate Statistical Analysis.Springer, 2005, pp325

* Sharma, S.: Applied Multivariate Techniques. John Wiley&Sons. Inc.,1996, pp185

Table4.1: The Code Numbers of the Variables

Good-Bad	A good or a bad customer depending on his/her credibility
1	Good
2	Bad
Customer type(3)	By which campaign the customers are included in the portfolio
1	B (By direct marketing)
2	C (Letter1)
3	D (Letter 2)
4	G (The employees working for other group institutitions
5	J (Private customers)
6	M (Mail Order)
7	N (Normal)
8	O (School campaigns)
9	R (Shopping Center)
10	T (Collective arrangements)
Age(1)	Customer's age.
1	Between18-25
2	Between26-33
3	Between 34-41
4	Between 42-49
5	Between 50-57
6	Bigger than 57
Gender(2)	
1	Male
2	Female
Car Ownership(6)	
1	Yes
2	No
Salary(4)	
1	0-700 YTL
2	701-1000 YTL
3	1001-1800 YTL
4	1801-3000 YTL
5	Bigger than3000 YTL
Net Monthly Income(5)	
1	0-700 YTL
2	701-1000 YTL
3	1001-1800 YTL
4	1801-3000 YTL
5	Bigger than 3000 YTL
Other Credit Card Ownership(7)	
1	1
2	2
3	None
Employment Type(8)	
1	Other
2	Retired
3	House wife
4	Public sector employee
5	Entrepreneur
6	Private sector employee
7	Student

4.1 Applying Discriminant Analysis

By using **SPSS for Windows 11.5**; 870 of the customers are qualified as "good ones" whereas 552 of the customers are "bad ones" as given in the Table 4.1.1 below:

Table 4.1.1: The Distribution of Good and Bad Customers

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	GOOD	870	61,2	61,2	61,2
	BAD	552	38,8	38,8	100,0
	Total	1422	100,0	100,0	

For the good customers and the bad customers the Gini type of heterogeneity coefficient results are given in the following table:

Table 4.1.2: Gini Type of Heterogeneity Scores after Classification by Discriminant Analysis

Groups	1	2	3	4
Good	0,641	0,563	0,457	0,824
Bad	0,63	0,614	0,453	0,6912
Total	0,64	0,67	0,458	0,9672
Average After classification	0,6355	0,5885	0,455	0,7576
Groups	5	6	7	8
Good	0,7878	0,621	0,53	0,418
Bad	0,6928	0,71	0,594	0,36
Total	0,9848	0,655	0,56	0,36
Average After classification	0,7403	0,6655	0,562	0,389

Here the group numbers are
1 for age; **2** for gender ; **3** for customer type ; **4** for salary; **5** for net monthly income; **6** for car ownership ; **7** for other credit card ownership and **8** for employment type.
Similarly, the entropy/maximum entropy ratios are in the table below;

Table 4.1.2: Entropy Type of Heterogeneity Scores after Classification by Discrimination Analysis

Groups	1	2	3	4
Good	0,876	0,339	0,763	0,961
Bad	0,868	0,511	0,754	0,888
Total	0,872	0,658	0,763	1
Average After classification	0,872	0,425	0,7585	0,9245
Groups	5	6	7	8
Good	0,944	0,533	0,608	0,67
Bad	0,892	0,741	0,69	0,606
Total	1	0,622	0,652	0,602
Average After classification	0,918	0,637	0,649	0,638

Here it should be noted that such a ratio has been proposed because the maximum of the entropy of a distribution depends on the number of groups. Therefore it seemed more logical to use a kind of index or a ratio to compare with the results of Gini coefficients directly.
The "total" columns of both of the tables show either the Gini type of coefficients or entropy/max entropy ratios for the whole observations. This column is added to measure the differences in entropies or Gini coefficients after the classification. For example for the first column of the Table 4.1.2, we can say that for the first variable the Gini type of heterogeneity coefficient is 0,64 before

classification. After the classification we obtain two groups (good and bad customers) whose heterogeneity coefficients are 0,641 and 0,63 and their average is 0,6355. So on the average there is a reduction of 0,641-0,6355 in the heterogeneity.

The first thing to note is that there is a certain parallelism between the the information given by the entropy and Gini type of heterogeneity scores although the scales are totally different. For example if any one of the coefficient reflects that there is a reduction in the degree of heterogeneity after classification the other gives similar information. So there is no contradiction between the information one measure gives and the one that the other gives.

The second thing to note is that, the classification procedure is not successful to decrease the level of heterogeneity for all classes.

The third thing to add (may be the most important result of all) is that the most dramatically reductions in heterogeneity scores are achieved in two of the three discriminating factors (namely the salary and net monthly income variables). May be this result shows the efficiency of such a classification procedure. Because the most serious reductions in heterogeneity measures (either by using entropy type of or Gini type of heterogeneity measures) are observed in these discriminating factors or classes.

4.2 Applying Cluster Analysis

As a second method of classification, we applied data a two step cluster analysis by SPSS for Windows 11.5; 663 of the customers are qualified as "good ones" whereas 759 of the customers are "bad ones" as given in the Table 4.2.1 below The results are given below;

Table 4.2.1: The Distribution of Good and Bad Customers

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	GOOD	663	46,6	46,6	46,6
	BAD	759	53,4	53,4	100,0
	Total	1422	100,0	100,0	

Table 4.2.2: Gini Type of Heterogeneity Scores after Classification by Cluster Analysis

Groups	1	2	3	4
Good	0,59	1	0,514	0,93
Bad	0,67	0,68	0,28	0,9
Total	0,64	0,68	0,46	0,97
Average After Classification	0,63	0,84	0,397	0,915
Groups	5	6	7	8
Good	0,93	1	0,51	0,39
Bad	0,92	0,83	0,61	0,39
Total	0,98	0,66	0,55	0,39
Average After Classification	0,925	0,915	0,56	0,39

Table 4.2.3: Entropy Type of Heterogeneity Scores after Classification by Cluster Analysis

Groups	1	2	3	4
Good	0,837	1	0,808	0,991
Bad	0,891	0,675	0,476	0,987
Total	0,872	0,412	0,763	1
Average After Classification	0,864	0,8375	0,642	0,989
Groups	5	6	7	8
Good	0,996	1	0,563	0,627
Bad	0,994	0,91	0,722	0,648
Total	1	0,62	0,652	0,648
Average After Classification	0,995	0,955	0,6425	0,6375

From **Table 4.2.2**; it can be seen that similar results are obtained in cluster analysis.

For the 4th variable the Gini type of heterogeneity coefficient is 0,97 before classification. After the classification the heterogeneity coefficients for the good and bad groups are 0,93 and 0,90 respectively. Their average is 0,915. So on the average there is a heterogeneity reduction with a magnitude of 0,055 (0,97-0,915 =0,055) For the 5th variable the Gini type of heterogeneity coefficient is 0,98 before classification. After the classification we obtain the scores fort he two groups as 0,93 and 0,92. So on the average there is a reduction of 0,98-0,925 in the heterogeneity

From **Table 4.2.3** ; similar results are obtained in cluster analysis. For the 4th variable the entropy type of heterogeneity coefficient is 1 before classification. After the classification we obtain two coefficients as 0,991 and 0,987 .This means that on the average there is a reduction of 0,991-0,987 in the heterogeneity for the 5th variable the entropy type of heterogeneity coefficient is 1 before classification. We obtain two coefficients fort he groups as e 0,996 and 0,994 and their average is 0,995. The reduction in the degree of heterogeneity is 0,005 (1-0,995=0,005) Of course this result is not as impressive as the previous results.

5. CONCLUSION

When the random variable/variables is/are qualitative, it is impossible to calculate the mean and the variance of any distribution.. Therefore it is necessary to use entropy as a measure of dispersion.

While the lower limit of entropy is zero, the upper limit depends on the number of groups of a distribution. The more the number of groups, the bigger the entropy. For this reason we used a measure of heterogeneity by using the ratio "entropy/maximum entropy".

As a second measure of heterogeneity, we adopted the methodology behind Gini index, by redefining the variables. Again by dividing the area under the curve to the total area which corresponds to the case of maximum heterogeneity, we obtained a second kind of heterogeneity index.

By applying two classification procedures to the data of a certain banking institution in Turkey, we aimed to obtain more homogeneous groups after classification. But since the data at hand is multivariate, some variables (groups) become more homogeneous while some others do not. In discrimination analysis, the discriminatory factors are chosen to the first, fourth and fifth variables. Therefore the reduction in heterogeneity scores due to classification especially for the fifth and sixth groups seemed meaningful.

The results related to heterogeneity scores of groups obtained by clustering are parallel to those of discrimination analysis although some poor results are obtained.

REFERENCES

1. Everitt; B.S.; Dunn ;E.; Applied Multivariate Data Analysis; Arnold; Second Edition; 2000
2. Hardle; W.;Simar; L.;Applied Multivariate Statistical Analysis;Springer, 2005
3. İpek; M.;Betimsel İstatistik; İpek; M.; Beta Yayınevi; 1991
4. Khinchin;A.Ya.; Mathematical Foundations of Information Theory ; Dover Publications; 1957
5. Press; S.J.; Applied Multivariate Analysis Using Bayesian and Frequentist Methods of Inference; Dover Publications; Second Edition; 2005
6. Reza ;F. M.;An Introduction to Information Theory; Dover Publications; 1994
7. Sharma;S.; Applied Multivariate Techniques; John Wiley&Sons;1996
8. SPSS for Windows 11.5 Help Manual
9. Turangil; N ; Kredi Skorlamada Kullanılan yöntemler ve Uygulamaları ;. Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü İstatistik Anabilim Dalı'nda Hazırlanan Yüksek Lisans Tezi; 2006
10. http://en.wikipedia.org/wiki/Gini_coefficient
11. http://mathworld.wolfram.com/GiniCoefficient.html