Selection of the K-MOST Important Attributes

Mohamed H. Farrag[1], **Maha M. Hana**[2], **Mona M. Nasr**[2]
[1]Theodor Bilharz Research Institute, Ministry of Scientific Research, Cairo, Egypt
eng.m.hassan@gmail.com
[2]Faculty of Computers and Information, Helwan University, Cairo, Egypt
mahana_eg@yahoo.com, m.nasr@helwan.edu.eg

Abstract

Organizations built their customer information data warehouse aiming to enhance the process of customer services which depends on different data mining techniques. Most of data mining techniques face a common problem which is the most important attribute to set as start node to begin with. To overcome this problem K-MIAS is a proposed methodology to select the K-Most important attributes that distinguish different customer types. K-MIAS methodology consists of three phases. The first phase is data preparation which prepares data for computing calculations. The second phase is K-MIAS algorithm which ranks the quantification levels for each attributes with respect to all attributes to select the K-Most important attributes while the third phase is to visualize data which helps for better data understanding and clarifying the results. In this paper, K-MIAS methodology is tested by a dataset consist of 1000 instants for trainee's questionnaire. K-MIAS methodology selects the K-Most important attribute successfully with interesting remarks and findings.

Keywords: *Data Mining, Customer Relationship Management, Customer Service.*

1. Introduction

Many organizations use different modern data analysis methods to support customer services process. Data mining involves the automated analysis of data to produce useful knowledge in a highly summarized form. Thus, Data mining is very useful in market segmentation, customer profiling, risk analysis, and other applications. It helps to enhance all customers' interactions in all the three stages of customer life cycle as acquiring customers, increasing the value of the customers and retaining good customers. An important role of data mining in different business process is classification process which aims to predict future customer behavior through classifying recent customer data. The most common tools for classification models are Neural Network, Decision Tree structure and if then – else rules.

Most of them face a common problem which is to identify the most important attributes to begin with. The current work addresses attribute selection by proposing an algorithm that helps to detect the K-Most important attributes that distinguish customers' clusters to enhance customer services process. Also, it helps to detect the most significant attributes applied in different classification data mining models. The proposed K-MIAS Methodology is a simple general methodology based on naïve statistical calculations methods to extract the most important attributes for the target dataset and helps to detect the appropriate value for selected attributes K. K-MIAS methodology supported by visualization aid to help users to examine the opportunities effect of different K values. Also, helps to find relation and the significant difference between the dataset attributes.

This paper is organized as follows: section two discusses related work, section three demonstrates the proposed methodology and section four explains the experiment while conclusion and future work are in the last section.

2. Related Work

Knowledge Discovery in Database (KDD) is the non-trivial process of identifying valid, potentially useful, and ultimately understandable patterns in data. The knowledge discovery process comprises six phases as Data selection, Data cleansing, Enrichment, Data transformation or encoding, Appling data mining tools and Displaying of the discovered information.

Customer Relationship Management (CRM) is defined as: "Communication with the customer, developing products and services that customers need, and selling or supporting them in manners that customer's desire"[1]. CRM is an Enterprise approach to understanding and influencing customer behavior through meaningful communications in order to improve customer acquisition, customer retention, customer loyalty, and customer profitability to build a profitable long term relationship with the customers in the field of marketing, sales, customer services and support.

The CRM framework classified in two classifications the first one is ***operational classification*** which is according to the business automation. While the second classification is the ***analytical classification*** which is according to the customer characteristics and behaviors to help an organization to effectively allocate resources to the most profitable group of customers or the target customers.

The CRM framework has four dimensions the first one is ***customer identification*** which concern with target customer analysis and customer segmentation, second is ***customer attraction*** which concern with direct marketing, the third is ***customer retention*** which concern with one to one marketing and loyalty programs while ***customer development*** is the fourth one concern with customer lifetime value analysis, up cross selling and market basket analysis.

Organization need to discover the hidden knowledge in the stored data to use it to acquire and retain potential customers and maximize customer return value that could be by using data mining tools which could help the organization to better discriminate and effectively allocate resources to the most profitable group of customers. Also, Data Mining is important tool to transform customer's data into meaning patterns to help in predicting and distinguishing different customer clusters. ***Data Mining*** is defined from different views and scopes. The first definition is "A branch of the applied informatics, which allows us to sift through large amounts of structured or unstructured data in attempt to find hidden patterns and/or rules". The second definition is "The data mining methods as tools for searching databases with special algorithms to identify general patterns which can be used in the classification of the individual observations and making predictions". Another definition is "Data mining is the search for valuable information in large volumes of data" [1].

Data mining has several roles in business process. First role is ***association*** which aims to establishing relationships between items which exist together in a given record. Second role is ***classification*** which use for building a model to predict future customer behaviors through classifying customer data into a number of predefined classes based on certain criteria. Third role is ***clustering*** which is to segment a heterogeneous population into a number of more homogenous clusters. Fourth role is ***forecasting*** which estimates the future value based on a record's patterns. It deals with continuously valued outcomes. Fifth role is ***regression*** which is a kind of statistical estimation technique used to map each data object to a real value provide prediction value. Sixth role is ***sequence discovery*** which is concerned with identification of associations or patterns over time. Seventh role is ***visualization*** which can be defined as the presentation of data so that users can view complex patterns.

Data Mining Algorithms the previous data mining model could be used through the next algorithms: ***Association Rule*** is defined as "a way to find interesting associations among large sets of data items". Also can be defined as "if /then statements that help to uncover relationships between seemingly unrelated data in a relational database or other information warehouse". An association rule has two parts, an antecedent (if) and a consequent (then). An antecedent is an item found in the data. A consequent is an item that is found in combination with the antecedent [5].

Decision Trees structure is a hierarchy of branches within branches that produces the characteristic inverted decision tree structure form. The nested hierarchy of branches is called a Decision Tree. Decision Tree structure is simple, although powerful form of multiple variable analyses. They provide unique capabilities to supplement, to complement, and to substitute a variety of data mining tools and techniques [4].

Genetic Algorithms (GAs) are adaptive heuristic search algorithm premised on the evolutionary ideas of natural selection and genetic. ***Artificial Neural Networks (ANN)*** is an information processing pattern. The key element of this model is the novel structure of the information processing system. It is composed of a large number of highly interconnected processing elements working in harmony to solve specific problems.

K-Nearest neighbour (K-NN) is a nonparametric method in that no parameters are estimated, output variables can either be interval variables in which case the K-NN algorithm is used for prediction while if the output variables are categorical, either nominal or ordinal.

Linear discriminate (LDA) / logistic regression (LR) they are widely used multivariate statistical methods for analysis of data with categorical outcome variables. Both of them are appropriate for the development of linear classification models. The difference between the two methods is LR makes no assumptions on the distribution of the explanatory data; LDA has been developed to give better results in the case when the normality assumptions are fulfilled.

Data Mining has two methods depends on the nature of use and nature of results. The first is ***Predictive Methods*** is depending on using some variables to predict unknown or future values of other variables. The second is ***Descriptive Methods*** is depending on find human-interpretable patterns that describe the data.

3. K-MIAS Methodology

The proposed methodology deals with multiple variable datasets to identify and rank the **K**- **M**ost **I**mportant **A**ttribute **S**election. It is a statistically based method featured with visualization aid.

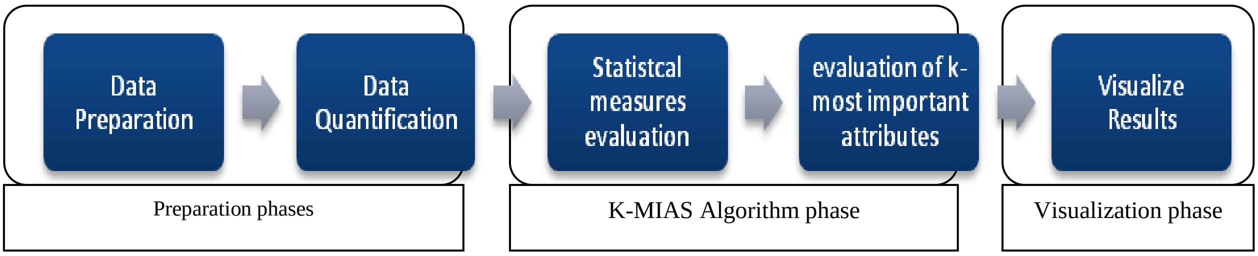

Figure (1): Methodology Steps

The methodology consists of three phases as shown in Figure (1).

Preparation Phase:

The first step is data preparation which aims to adopt the dataset values to be suitable for computer calculations. It numerates the dataset attribute, revises the data for odds values and missed values.

The second step is data quantification which aims to decrease numerical data variability and puts the data values into one quantification level format. Quantification levels are set according to the attribute minimum and maximum values, the density distribution for attribute values, its scope and nature.

K-MIAS Algorithm Phase:

The first step is a proposed statistical measures evaluation in which the data undergo a set of statistical calculations to generate and to produce new statistical parameters.

The second step is the evaluation of k-most important attributes. K-MIAS algorithm ranks the quantification levels for each attributes with respect to all attributes.

Visualization phase

The first step consists of two sets of charts. The first set visualizes results and the second set visualizes the dataset and intermediate computations as follows:

First set

- A chart displays the output ratio average ranked according to K value of all attributes and its quantification levels.
- A chart displays the output ratio average ranked according to the values for all attributes and its quantification levels. This chart shows the difference between the selected attributes, also the difference between them and the other attributes.

Second set

- A chart displays the output ratio average for all attributes and its quantification levels. This chart shows the relation between the numbers of quantification levels each attribute and the selected attributes.
- A chart displays the distribution for all instants to the selected clusters for all attributes and its quantification levels. This chart shows the relation between the number of instants for quantification level for each attribute and the selected attributes.

K-MIAS algorithm

This section describes the algorithm to extract the K-Most significant important attributes. It starts with specifying the used notations, functions and K-MIAS algorithm itself as shown in Figures (2,3,4) respectively. This section ends with demonstrating examples illustrated in Table (1) and Table (2).

Algorithm Notations

C_i: i^{th} cluster where i=1...c

C_s: denotes the selected cluster which its behavior is to be mined.

A_k: k^{th} attribute where k=1..A

QL_m: m^{th} quantification level for each attribute where m=1...QL

A_k-QL_m-C_i: denotes an instance of i^{th} cluster in the m^{th} quantification level in k^{th} attribute.

SMC: denotes the sum of all number of instants for all quantification levels for all attributes to each cluster

SMCQL: denotes the sum of all number of instants for each quantification level for each attribute for each cluster

A_k-QLm-C_iC_sR: denotes the ratio between A_k-QL_m-C_i and *SMC* to the selected cluster C_s

A_k-QL_m-CsRC_i: denotes the ratio between members of **A_k-QL_m-C_i** to sum of all members for all clusters except the selected cluster

$OUTRATIO_{avg}$: denotes average between A_k-QL_m-C_iC_sR and A_k-QL_m-C_sRC_i

$OUTRATIO_{avg\_rank}$: denotes ranked attribute quantification levels to k most important attributes

RAk-QLm: denotes the ranked attributes quantification level after remove redundancy.

Figure (2): K-MIAS Algorithm Notations

Algorithm Functions

Member(x) = no. of members in set (x)

Avg (X, Y)= arithmetic average of (X & Y)

SM(x) =arithmetic sum of members in set (x)

Rank (X) = sort elements of (X) & remove redundant value

Figure (3): K-MIAS Algorithm functions

Algorithm Steps

Input: C_i, A_k, QL_m, K, C_s

Output: K-most important attributes.

Begin:

Step 1:

C_i=member(C);

A_k =member (A);

QL_m =member (QL);

Step 2:

$$\{\forall A\kappa\{\forall Qlm\{\forall C\iota, A\kappa - Qlm - C\iota = member(A\kappa - Qlm - C\iota) | \kappa = 1 \ldots A, m = 1 \ldots QL, \iota = 1 \ldots C \text{ where } A\kappa \in A, QLm \in QL \ \& \ C\iota \in C\}\}\}$$

Step 3:

$$\{\forall A\kappa\{\forall Qlm\{\forall C\iota, SMC = SM(A\kappa - Qlm - C\iota) | \iota = 1 \ldots C, m = 1 \ldots QL, \kappa = 1 \ldots A \text{ where } A\kappa \in A, QLm \in QL \ \& \ C\iota \in C\}\}\}$$

Step 4:

$$\{\forall A\kappa\{\forall Qlm\{\forall C\iota, SMCQL = member(A\kappa - Qlm - C\iota) + member(A\kappa - Qlm - Cj) | \iota \neq j, \iota, j = 1,2 \ldots C \ \& C\iota, Cj \in C\}\}\}$$

Step 5:

$$\{\forall A\kappa\{\forall Qlm\left\{\forall C\iota, A\kappa - Qlm - C\iota CsR = \frac{member(A\kappa - Qlm - C\iota)}{member(Cs)} | \iota \neq s, \iota = 1,2 \ldots C \ \& C\iota, Cs \in C\right\}\}\}$$

Step 6:

$$\{\forall A\kappa\{\forall Qlm\left\{\forall C\iota, A\kappa - Qlm - CsRC\iota = \frac{member(A\kappa - Qlm - Cs)}{SMCQL - member(A\kappa - Qlm - Cs)} | \iota \neq s, \iota = 1,2 \ldots C \ \& C\iota, Cs \in C\right\}\}\}$$

Step 7:

$$\left\{\forall A\kappa\{\forall Qlm\{\forall C\iota, OUTRATIOavg = avg(A\kappa - Qlm - C\iota CsR, A\kappa - Qlm - CsRC\iota) | \iota \neq s, \iota = 1,2 \ldots C \ \& C\iota, Cs \in C\}\}\right\}$$

Step 8:

$OUTRATIO_{avg\_rank}$ =Rank ($OUTRATIO_{avg}$)

Step 9:

Output K-First RA_k-QL_m

End

Figure (4): K-MIAS Algorithm Steps

Table (1): Dataset attributes and its quantification levels

Attribute		Clusters			*SMCQL*	*Ak-QLm-CiCsR*	*Ak-QLm-CsRCi*	*OUTRATIOavg*
		C1	**C2**	**C3**				
Name	**AQL**	***SMC=85***	***SMC=67***	***SMC=48***		**(16/85)**	***(16/(28-16))***	
A1	**A1QL1**	***Ak-QLm-Ci =16***	***Ak-QLm-Ci =8***	***Ak-QLm-Ci =4***	28	**0.188**	**1.333**	**0.76**
	A1QL2	**8**	**4**	**6**	18	**0.094**	**0.800**	**0.45**
	A1QL3	**23**	**19**	**12**	54	**0.271**	**0.742**	**0.51**
A2	**A2QL1**	**9**	**22**	**13**	44	**0.106**	**0.257**	**0.18**
	A2QL1	**29**	**12**	**25**	66	**0.341**	**0.784**	**0.56**

Table (2): K-MIAS algorithm ranked attributes quantification level according to $OUTRATIO_{avg\_rank}$

Attribute		Clusters			*SMCQL*	*Ak-QLm-CiCsR*	*Ak-QLm-CsRCi*	*OUTRATIOavg_rank*
		C1	**C2**	**C3**				
Name	**AQL**	***SMC=85***	***SMC=67***	***SMC=48***		**(16/85)**	***(16/(28-16))***	
A1	**A1QL1**	***Ak-QLm-Ci =16***	***Ak-QLm-Ci =8***	***Ak-QLm-Ci =4***	28	**0.188**	**1.333**	**0.76**
A2	**A2QL1**	**29**	**12**	**25**	66	**0.341**	**0.784**	**0.56**

Table (1) and (2) show an example of K-MIAS methodology results applying on synthetic dataset that consists of 100 instants in three clusters with two attributes and a total of five quantification levels as explained below:

<u>Input for *K-MIAS* is:</u>
Dataset has three clusters which are (“C1”, “C2”, “C3”). It has two qualitative attributes (“A1”, “A2) and five quantification levels. Let ***K*** equal two and ***Cs*** is”C1”.

<u>Step1</u>
It is identifies the numbers of all clusters, all attributes and all quantification levels.
For example, dataset in Table 2 has the following: C for “C1” are 85 instants, the member for attribute “A1” is 100, members of “A1 (A1qL1)” quantification level is 28

<u>Step2</u>
K-MIAS algorithm identifies the number of instants for each cluster for each attribute quantification level to generate $\boldsymbol{A_k\text{-}QL_m\text{-}C_i}$.
For example, the number of instants for the attribute “A1” to the first quantification level "A_1QL_1" for “C1” is 16 instants and “C2” is 8 instants while “C3” is 4 instants.

<u>Step3</u>
K-MIAS algorithm identifies the sum for all number of instants for all quantification level for all attributes to each cluster, ***SMC =85***.
For example, the total number of instants for all attributes quantification levels for the cluster “C1” is 85 instants and 67 for “C2” cluster while “C3” cluster is 48.

<u>Step4</u>
K-MIAS algorithm identifies the sum for all instants for all clusters for a specific quantification level to generate ***SMCQL*** and repeated for all quantification levels in each attribute.
For example, the total numbers of instants to the first quantification level “A1QL1" for all clusters is 28 instants.

<u>Step5</u>

K-MIAS algorithm identifies the ratio between A_k-QL_m-C_i and *SMC* for ***Cs*** to generate $\boldsymbol{A_k\text{-}QL_m\text{-}C_iC_sR}$
For example, the ratio for the first quantification level “A1QL1" for the ***Cs*** is 16/85. This step is to measure the power of this quantification level among the rest of quantification levels for the tested cluster ***Cs***.

<u>Step6</u>
K-MIAS algorithm identifies the ratio between A_k-QL_m-C_i and (*SMCQL* - A_k-QL_m-C_i) for ***Cs*** to generate $\mathbf{A_k\text{-}QL_m\text{-}C_sRC_i}$
For example, the ratio for the first quantification level "A1QL1" for ***Cs*** to all other clusters is 16/ (28-16). This step is to measure the power of this quantification level for the selected cluster ***Cs*** among its values to the rest of other clusters in same quantification level.

Step7

K-MIAS algorithm identifies the average ratio between A_k-QL_m-C_iC_sR and A_k-QL_m-C_sRC_i to generate ***OUTRATIO$_{avg}$***

For example, the *OUTRATIO$_{avg}$* for first quantification level "A1QL1" is 0.188 and 1.333 equal 0.76.

Step8 & step9

K-MIAS algorithm ranks the output according to values of OUTRATIOavg as ***OUTRATIO$_{avg_rank}$*** and removes the redundancy to generate ***RA_k-QL_m*** to insure that the K most important attributes does not belongs to the same attribute by selecting the highest quantification level in each attribute.

Output as k=2 is Quantification Level "A1Ql1" for the attribute "A1", Quantification level "A2Ql1" for the attribute "A2" as shown in Table 2.

4. Experiment

Data Set Description

The methodology uses a dataset that has been obtained from one of private training center in the field of computer and language education. This center faces some problems to find the most important attributes that distinguish specific types of customers who preferred to join specific track from huge data stored as free-text trainee's questionnaire to support customer services process. The dataset consists of 1000 instants that has been transformed from original free-text trainee's questionnaire into database tables and excel spreadsheet formats. The dataset are divided into three clusters of customers. Clustering is based on measuring the interest of customers to join specific track of courses. The first cluster is "Interested cluster" has 435 instants while the second cluster has 170 instants for "maybe cluster". The last cluster has 312 instants for "uninterested cluster" and 83 instants have missed values. The data set has seven qualitative attributes that are different in nature and scope. The attributes are age, gender, education level, media, education type, employment and location as shown in Table (3).

Table (3): Example of Attributes Status

#	**Attributes**	**Status**
1	Age	qualitative
2	Gender	qualitative
3	Education Level	qualitative

4.1 Data preparation phase

This phase aims to detect and classify dataset attributes, then assign observations values from dataset to each attribute and clear dataset by removing 83 instants with missed values.

4.2 Data quantification phase

This phase divides these attributes into ranges according to minimum and maximum values and data distribution values.

Table (4): Example of Attribute Quantification Levels Range

Attribute		CLUSTERS		
Name	**AQL**	**Interested (435)**	**May be (170)**	**Uninterested (312)**
Age	15-20 (230)	102	48	80
	21-25 (278)	151	66	61
	26-30 (240)	137	16	87
	31-40 (142)	37	36	69
	>40 (27)	8	4	15
Gender	Male (345)	262	15	68
	Female (572)	173	155	244
Education type	CS & Engineering	109	12	30
	Commerce	46	17	49
	Literary education	196	106	135
	Other	84	35	98

The algorithm quantify the numerical attributes by dividing them 2 to 4 quantification levels according to maximum instant value, minimum instant value and the density of instants distribution. Each quantification level must include all instants between the minimum value and the maximum value for this quantification level range. Insure that no intersection between these quantification levels.

For example, a numerical attribute called “ABC” the minimum instant value is 4 and the maximum instant value is 72 while the density of instants values distribution displayed as shown in Figure (5).

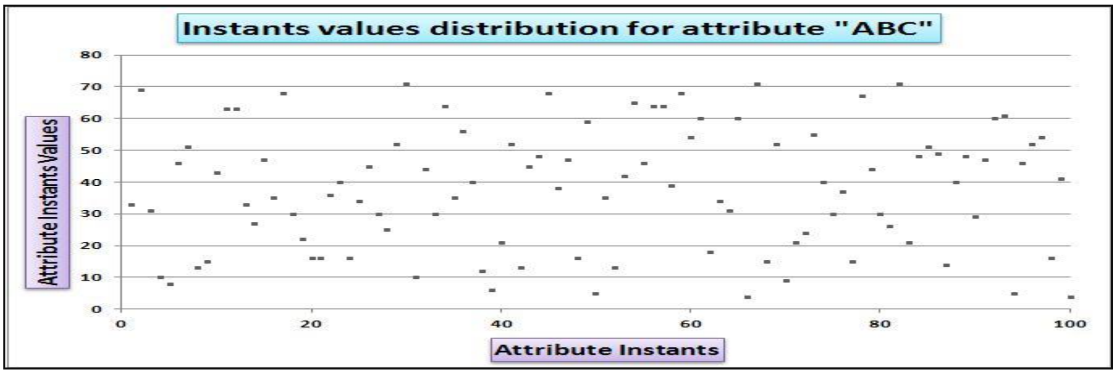

Figure (5): Examples of instants values distribution for attribute “ABC”

The algorithm divides the attribute “ABC” into four quantification levels according to the minimum, maximum values and the density of instants values distribution using rank function as shown in Table (5).

Table (5): Example of dividing numeric Attribute into Qualitative Status

#	Attribute	AQL	AQL Range	No. Of Instants
1	ABC	AQL1	... <= 24	**28**
		AQL2	25 ... <= 40	**26**
		AQL3	41 ... <= 52	**23**
		AQL4	>= 53	**23**

4.3 K-MIAS algorithm phase

The following are demonstrative steps for K-MIAS algorithm.

Input for *K-MIAS* is:
Dataset has three clusters which are (“Interested customers”, “May be customers”, “Uninterested customers”). It has seven qualitative attributes (“age”, “gender”, “education level”, “media”, “education type”, “employment”, and “location”) and Twenty four quantification levels. Let ***K*** equal three and ***Cs*** is “Interested Cluster”.

Step1
It is identifies the numbers of all clusters, all attributes and all quantification levels.
For example, dataset in Table 4 has the following: C for “may be cluster” are 170 instants, the member for attribute “age” is 913, members of “age (15-20)” quantification level is 230.

Step2
K-MIAS algorithm identifies the number of instants for each quantification level in each cluster to generate $\boldsymbol{A_k\text{-}QL_m\text{-}C_i}$.
For example, The number of instants for the attribute “age” to the first quantification level “15 years -20 years” A_1QL_1 for “Interested Cluster” is 102 instants and “May be cluster” is 48 instants while “uninterested cluster” is 80 instants.

Step3
K-MIAS algorithm identifies the sum for all number of instants for all quantification level for all attributes to each cluster, ***SMC =917***.
For example, the total number of instants for all attributes quantification levels for the cluster “Interested” is 435 instants and 170 for “May be” cluster while “uninterested” cluster is 312.

Step4
K-MIAS algorithm identifies the sum for all instants for all clusters for a specific quantification levels to generate ***SMCQL*** and repeated for all quantification levels in each attribute.
For example, the total number of instants to the first quantification level “15 years -20 years” A_1QL_1 for all clusters is 230 instants.

Step5
K-MIAS algorithm identifies the ratio between A_k-QL_m-C_i and *SMC* for ***Cs*** to generate ***A_k-QL_m-C_iC_sR***
For example, the ratio for the first quantification level “15 years -20 years” A1QL1 for the ***Cs*** is 102/435.

Step6
K-MIAS algorithm identifies the ratio between A_k-QL_m-C_i and (*SMCQL* - A_k-QL_m-C_i) for ***Cs*** to generate **A_k-QL_m-C_sRC_i**
For example, the ratio for the first quantification level “15 years -20 years” A_1QL_1 for ***Cs*** to all other clusters is 102/(230-102)

Step7
K-MIAS algorithm identifies the average ratio between A_k-QL_m-C_iC_sR and A_k-QL_m-C_sRC_i to generate ***$OUTRATIO_{avg}$***
For example, the *$OUTRATIO_{avg}$* for first quantification level “15 years -20 years” A_1QL_1 is 23.45 and 79.69 equal 51.57

Step8 & step9
K-MIAS algorithm ranks the output according to values of OUTRATIOavg as ***$OUTRATIO_{avg\_rank}$*** and removes the redundancy to generate ***RA_k-QL_m***
Output as k=3 is Quantification Level "Male" for the attribute "Gender", Quantification level "Cs & Engineering " for the attribute "Education type" and quantification level "Friends" for the attribute "Media".

Table (6): Example of K-MIAS algorithm Results

Attribute				Clusters			*SMCQL*	*Ak-QLm-CiCsR (102/435)*	*Ak-QLm-CsRCi (102/(230-102))*	*OUTRATIOavg*
#	Name	AQL	AQL Range	Interested *SMC=435*	May be *SMC=170*	Uninterested *SMC=312*				
A1	**Age**	**A1QL 1**	**15-20**	***Ak-QLm-Ci* =102**	***Ak-QLm-Ci* =48**	***Ak-QLm-Ci* =80**	230	0.23	0.80	51.57%
		A1QL 2	**21-25**	**151**	**66**	**61**	278	0.35	1.19	76.81%
		A1QL 3	**26-30**	**137**	**16**	**87**	240	0.31	1.33	82.25%
		A1QL 4	**31-40**	**37**	**36**	**69**	142	0.09	0.35	21.87%
		A1QL 5	**>40**	**8**	**4**	**15**	27	0.02	0.42	21.97%

4.4 Visualization phase

Visualization phase aims to help to analyze data and interpreting the output results represented into two sets: The first set is shown in Figure (6), (7).

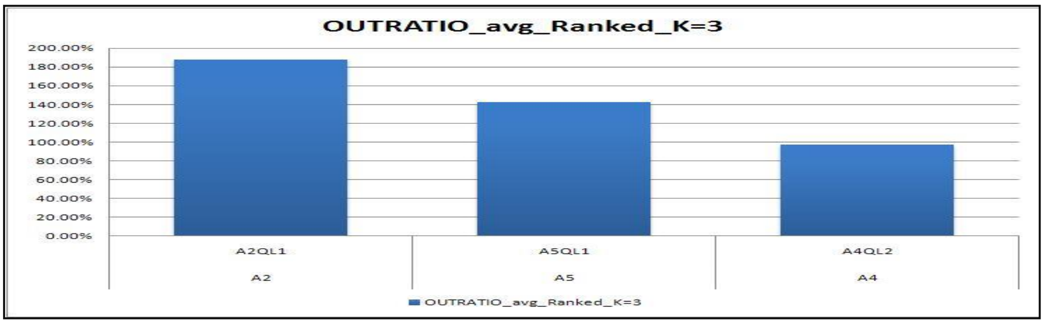

Figure (6): The Output Ratio Average Ranked According to Select Number of K for All Attributes and its Quantification Levels

Figure (6) shows the selected three most important attributes by the K-MIAS Methodology which are quantification Level "Male" for the attribute "Gender", quantification level "Cs & Engineering " for the attribute "Education type" and quantification level "Friends" for the attribute "Media".

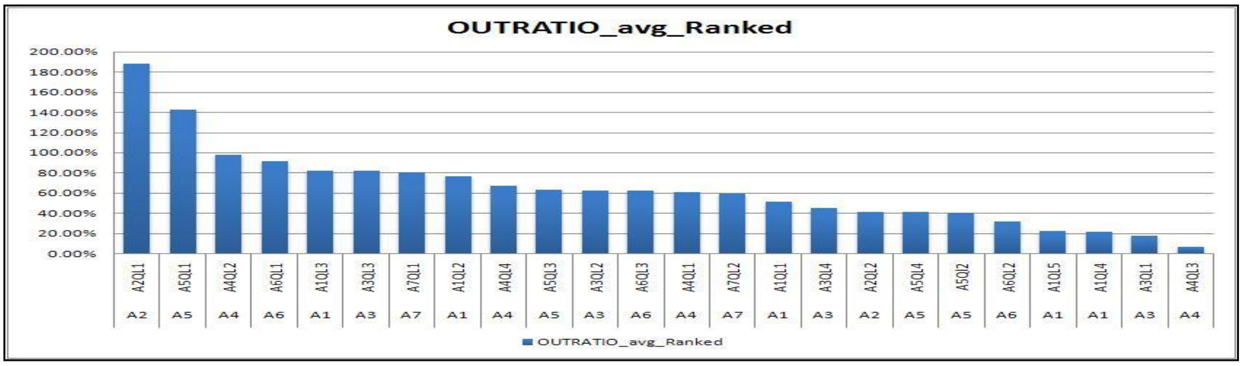

Figure (7): The Output Ratio Average Ranked According to the Values for All Attributes and its Quantification Levels

Figure (7) shows the rank for all quantification levels for all attributes to find the difference between the selected attributes, also the difference between them and the other attributes. It can be inferred that K equal three is an appropriate value because it is notable that there is a significant difference between the first and the second quantification level and also between the second and third one, while there is a slight difference between the third and fourth one and for the rest pairs.

The second set is shown in Figure (8), (9).

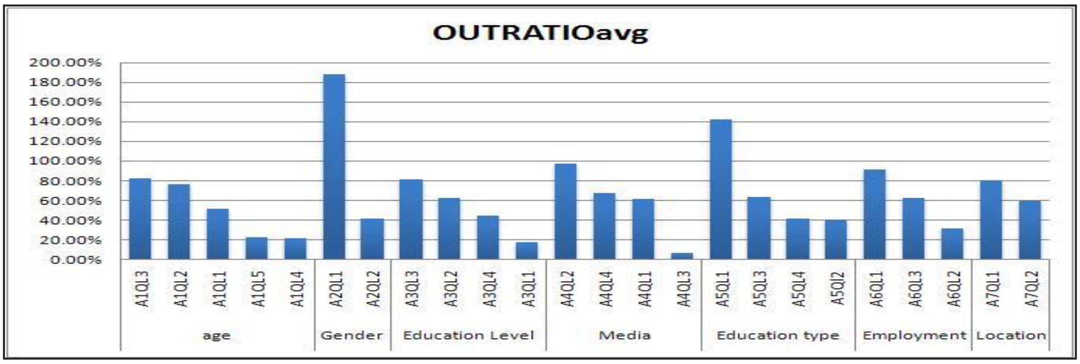

Figure (8): The Output Ratio Average for All Attributes and its Quantification Levels

Figure (8) shows two remarks. The first is the ranking of each quantification level in an attribute and the second remarks shows the distinctive power among all quantification levels on the same scale, helps to find the relation between the numbers of quantification levels for each attribute and the selected attributes. It is notable that the most important attributes does not depend on the number of quantification levels for each attribute as represented for the attribute "Age" which has the highest number of quantification levels reaches five is not selected K-most important attribute.

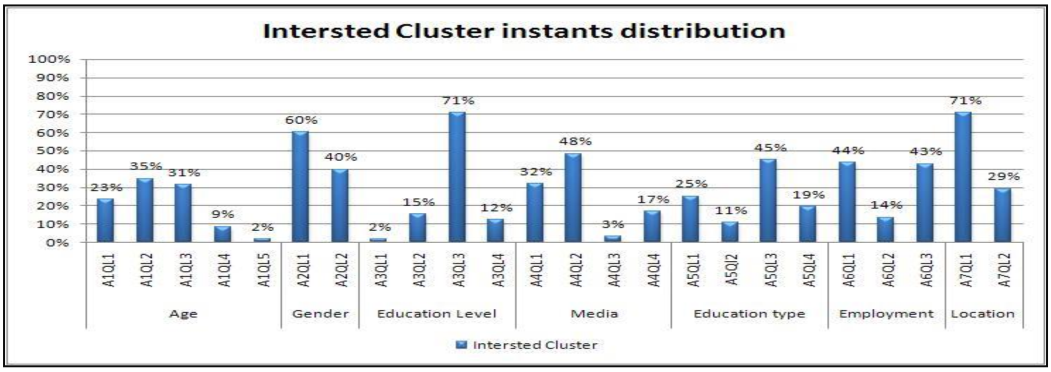

Figure (9): The Distribution for All instants to the Selected Clusters for All Attributes and its Quantification Levels

Figure (9) helps to understand the distribution of number of instants according to the selected clusters "Interested Cluster" among all attributes and its quantification levels. This chart shows the relation between the number of instants for quantification level for each attribute and the selected attributes. It is notable that the most important attributes does not depend on the number of instants for quantification level in the attribute as shown it selects A_2Ql_1, A_5Ql_1 and A_4Ql_2. In spite of that both quantification level A_7ql_1 and A_3ql_3 have the heights instants percentage that reaches up to 71%, as shown Figure (9).

4.5 Results and Evaluation of K-MIAS methodology

K-MIAS Methodology extracts the most important quantification levels and its attributes for all dataset attribute where K=3 as Quantification Level "Male" for the attribute "Gender", Quantification level "Cs & Engineering " for the attribute "Education type" and quantification level "Friends" for the attribute "Media" as shown in Figure (6).

It can be inferred that K equal three is an appropriate value because there is a significant difference between the first and the second quantification level and also between the second and third one. While there is a slight difference between the third and fourth one and for the rest pairs as shown in Figure (7).

K-MIAS Methodology shows that the most important attribute does not depend on the number of instants for quantification level in the attribute or the number of quantification levels for each attribute as shown in Figure (9). It selects A_2Ql_1, A_5Ql_1 and $A_4Ql_{2.}$ In spite of that both quantification level A_7ql_1 and A_3ql_3 have the heights instants percentage that reaches up to 71%, as shown Figure (9). Also, the attribute "Age" which has the highest number of quantification levels reaches five is not selected K-most important attributes as shown in Figure (8). Therefore, it is inferred that neither the instants percentage nor the number of quantification levels has any effect on the K-MIAS algorithm.

5. Conclusion and Future Work

K-MIAS methodology is a new proposed methodology to select K-Most important attributes. It is a simple general methodology based on naive statistical calculations methods. K-MIAS is a simple methodology as it doesn’t utilize sophisticated techniques or excessive computations.

K-MIAS is a general methodology for two reasons. First, it doesn’t have any special requirements and second its implementation is flexible enough to adopt completely different data domains and different datasets structures.

One of the advantages of K-MIAS Methodology is that its results is independent on the number of instants for quantification level in an attribute or the number of quantification levels for each customer attribute. Another advantage of K-MIAS methodology is that visualizing the results enable user to examine the opportunities effect of different K values.

The accuracy and the capacity of K-MIAS seem to be adequate and accurate on this work, yet it needs to be proved after using different classification techniques on different datasets.

It is a suggested that original K-MIAS methodology be modified to give attributes and its quantification levels different weights. Then, compare the results to the original K-MIAS methodology. Also, it is suggested to test K-MIAS using different data domains and study its accuracy and capacity on different data mining techniques.

References

[1] Gramatikov, M., (2003), Data Mining Techniques and the Decision Making Process in the Bulgarian Public Administration, NISP Acee Conference, Bucharest, Romania.

[2] Srivastava, J. (2000 January), Data Mining for Customer Relationship Management, ACM SIGKDD Explorations Newsletter, Volume 1, No 2.

[3] Agrawal, D., (2007 November), Building Profitable Customer Relationships with Data Mining, CSI Research Journal of India.

[4] Jia-Lang Seng & T.C. Chen, (2010 December), An Analytic Approach to Select Data Mining for Business Decision, Taiwan Expert Systems with Applications, Volume 37, Pages 8042–8057.

[5] Giha, E. F., Singh, P.Y., & Ewe, T. H., (2006 June). Mining Generalized Customer Profiles, AIML 06 International Conference, Sharm El Sheikh, Egypt, Volume 6, Pages 141-147.

[6] Hsieh, N. & Chu, K., (2009), Enhancing Consumer Behavior Analysis by Data Mining Techniques, International Journal of Information and Management Sciences, Volume 20, No 1, Pages 39-53.

[7] İkizler, N., & Güvenir, H., (2001), Mining Interesting Rules in Bank Loans Data, Proceedings of the Tenth Turkish Symposium on Artificial Intelligence and Neural Networks, Pages 238-246.

[8] Bartok, J., Habala, O., Bednar, P., Gazak, M. & Hluchý, L., (2010) Data Mining and Integration for Predicting Significant Meteorological Phenomena, ICCS 2010 Procedia Computer Science Volume 1, No 1, Pages 37–46.

[9] Çiflikli, C., & Kahya-Özyirmidokuz, E., (2010 December), Implementing a data mining solution for enhancing carpet manufacturing productivity. Knowledge-Based Systems, Volume 23, No 8, Pages 783–788.

[10] Pauray S.M., (2010), Mining top-k frequent closed item sets over data streams using the sliding window model, Taiwan Expert Systems with Applications Volume 37, Pages 6968–6973.

[11] Kamrunnahar, M., & Urquidi-Macdonald, M., (2010 March), Prediction of corrosion behavior using neural network as a data Mining tool, Corrosion Science, Volume 52, No 3, Pages 669–677.

[12] Thanuja V., Venkateswarlu, B., & Anjaneyulu, G. S. G. N., (2011 June), Applications of Data Mining In Customer Relationship Management. Journal of Computer and Mathematical Sciences Volume 2, No 3, Pages 399-580.