Data Mining Application for the Performance of Indian Industries using Financial Ratios

R. Lakshmi Priya
Assistant Professor,
Department of Statistics
Dr. Ambedkar Govt. Arts College,
Vysarpadi Chennai, India

R. Chandarasekaran
Associate Professor and Head (Retd.)
Department of Statistics
Madras Christian College.
Chennai, India

ABSTRACT
This paper examines the effectiveness of Data Mining and classification techniques in detecting Indian Industrial Performance (IIP) using financial ratios and deals with identification of factors associated to IIP. In this research paper, industries are split into five major groups and analyzed using traditional multivariate analysis and data mining techniques based on fourteen financial ratios. The present work is also intended to analyze financial performances of the five groups to understand financial scenario and to assess their financial strengths to enable the decision makers for future planning. The dataset relates to 445 companies of five major industrial sectors from Indian corporate database. The dataset comprises of important financial ratios of 78 companies from cement, 115 companies from steel, 102 companies from plastic, 66 companies from leather and 84 companies from hardware and software industries. The time frame of the data pertaining to the present study is 2001-2010. The salient feature of this study is the application of *Factor Analysis*, *K-means clustering* and *Discriminant Analysis* as data mining tools to explore the hidden structures present in the dataset for each of the study periods. Factor analysis is applied for Data Reduction and extraction of the hidden structure in the original data set. The financial ratios are used to find initial and final groups by k-means clustering algorithm. A few outlier industries, which could not be classified to any of the group, are discarded, as some of the ratios possessed unusual values. Finally, to cross validate the final clusters obtained by *k-means algorithm*, *Discriminant Analysis* is used to identify the industries as belonging to **EP**-Class (Elevated Performance), **MP**-Class **(**Moderate Performance**)** and **SP**-Class (Stumpy Performance). The results of the present study indicate that *k-means clustering algorithm* and *Discriminant Analysis* can be used as a feasible tool to analyse large set of financial data.

Keywords
Data Mining, Financial ratios, Factor Analysis, k-means clustering, Discriminant Analysis and Classification

1. INTRODUCTION
In the robust world, growth of industrial sector in the semi developed countries like China, India, etc. depends significantly on major and minor industries. Industry is the segment of economy concerned with production of goods, price, marketing and other services. An industrial growth is based on various stages like, origin and development.

Any management must put in effort to extend the development stage till they develop new products through research. But regrettably, few industries fail even if they use study approaches. A business failure is unfortunate. Some of the industries may fail within the first year or second year of their life due to various reasons and other industries may develop, mature and perform much later. In this study, we carry out an in-depth examination of Indian Corporate database using financial ratios in order to identify their performance by Data Mining and Multivariate Statistical Analysis.

In this study, three Statistical Data Mining techniques are used on industrial data for dimensionality reduction of financial ratios and to assess performance of industries: Principal Component Analysis (PCA), k-means clustering technique and Multivariate Discriminant Analysis. The three methods are compared in terms of their predictive classification and accuracy. The input data consists mainly of financial ratios derived from financial statements of companies. The sample contains data from 447 companies from five Indian Industrial sectors. The paper is arranged as follows: Section 2 reviews relevant previous research work. Section 3 provides database and variables. Section 4 presents various research methodologies. Section 5 describes the results and discussion. Finally, section 6 presents the concluding remarks.

2. REVIEW OF LITERATURE
Financial ratios have been used as inputs for advanced statistical models to forecast many kinds of business events and identify financial performance and other characteristics, but the focus has been on testing statistics models to predict business failures and performance. Altman (1968) was the first person to conduct a pioneering work in the year 1968 using financial ratios. The first step was to bridge the gap between usual ratio analysis and the more accurate statistical techniques, and predict corporate bankruptcy. Specifically, a set of financial ratios were investigated in a the context of bankruptcy prediction wherein a multivariate discriminant analysis based methodology was employed. The results were encouraging, suggesting that bankruptcy prediction is an accurate forecaster of failure.

Deakin (1972) proposed an alternative model using a different set of financial ratios to predict business failure. He used dichotomous classification test to determine the error rates that would be experienced when firms were classified on the basis of their financial ratios. By developing a discriminant model, the results were compared with dichotomous classification. Altman *et.al.* (1968, 1984 and 1994) developed a multiple discriminant model to determine the credit worthiness of commercial loan applicants in a particular troubled industrial sector in France. Chandrasekaran *et. al.,* (2013) have graded companies that reflected the performance of companies based on certain financial ratios. They choose a set of forty one ratios in this study. Principal Component Analysis (PCA) was also applied to understand the global nature of financial ratios and reduce the dimensionality. However, the data reduction of number of ratios showed a great fall in the accuracy of discrimination.

Data Mining (DM) is an iterative process within which progress is defined by detection, either through automatic or manual models. DM is most useful in an exploratory analysis scenario in which there are no predetermined notations about what will constitute an interesting outcome (Kantardzic, 2001). The application of DM techniques for financial classification is a rich research area. In the recent years, studies were conducted on leading Indian Industries based on their financial ratios and using various Statistical Data Mining techniques (Chandrasekaran R, Manimannan G, and Lakshmi Priya, 2014). Without making any assumptions with regards to the number of groups or any other structural pattern in advance, the performance of companies are studied based on certain financial ratios using the concept of DM. The results from the study has promised only three meaningful classifications of a company that could be rated as Grade-HP (High Performance), Grade-MP (Moderate Performance and Grade-LP (Low Performance), where Grade-HP companies are superior to Grade-MP and Grade-LP companies, Grade-MP companies are superior to Grade-LP companies but inferior to Grade-HP companies in terms of mean values of certain financial ratios (Chandrasekaran R, Manimannan G, and Lakshmi Priya, 2014).

3. DATABASE AND VARIABLES

This section brings out the discussion of the database, the ratios selected and the Data Mining Techniques.

3.1 Database and Selection of Variables

The financial data published by *Capital Market* (Indian Corporate Database) was considered as the database. The data mainly consists of five major types of industries in India and under each type of industry, there are several companies. The data consists of financial ratios of companies for the time period of ten years from 2001 to 2010, around 445 companies from each industry type. Among the listed companies, number of companies varied over the study period owing to removal of those companies for which the required data are not available. In this study, 14 ratios are carefully chosen from among the many that had been used in previous studies *(Table 1)*. These 14 ratios are chosen to assess profitability, solvency, liquidity, and cash-equity ratio. The ratios selected in the present study are based on two main criteria, namely their popularity as evidenced by their frequent usage in the finance and accounting literature and that the ratios have been shown to perform well in previous studies.

4. RESEARCH METHODOLOGY

4.1 Data Mining Techniques

Although data mining is relatively a new term, the technology is not. Data Mining or Knowledge Discovery in Databases (KDD) is the process of discovering previously hitherto unknown and potentially useful information from the data in databases. In the present context, data mining techniques exhibits the patterns by applying the techniques namely, factor analysis, **k**-means clustering and discriminant function rules. As such KDD is an iterative process, which mainly consists of the steps in the following figure:

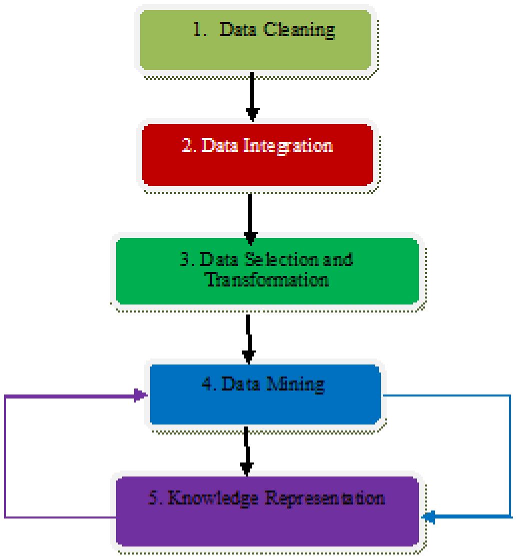

Figure 1. Data Mining and their Process

Of the above iterative process, Steps 4 and 5 are very important. If appropriate techniques are applied in Step 5, it provides potentially useful information that explains the hidden structure. This structure discovers knowledge that is represented visually to the user, which is the final phase of data mining.

4.2 Factor Analysis

Use of factor analysis, a multivariate statistical technique, increased during the past few decades in the fields of business related research. As the number of parameters to be considered in multivariate techniques increases, so does the need for increased knowledge of the structure and interrelationships of the variables. Factor analysis gives the tools for analyzing the structure of the interrelationships (correlations) among the large number of variables by defining sets of variables, mostly labeled, known as factors (Anderson, 1984). In the present study, factor analysis is initiated to uncover the patterns underlying financial ratio (*Table 1*). In factor extraction method the number of factors is decided based on the proportion of sample variance explained. Orthogonal rotations such as Varimax and Quartimax are used to measure the similarity of a variable with a factor by its factor loading (Everitt and Dunn, 2001; Hair, Black, Babin and Anderson, 2010).

Table 1. List of Financial parameters used in the Present study

Ratios	Description
DEB_EQU	Debt - Equity Ratio
LONG_TE	Long Term Debt-Equity Ratio
CURREN	Current Ratio
FIX_ASS	Fixed Assts
INVENTO	Inventory

DEBTORS	Debtors
INTERES	Interest
PBDITM	Profit Before Depreciation Interest Tax Margin
PBITM	Profit Before Interest Tax Margin
PBDTM	Profit Before Depreciation Tax Margin
CPM	Current Profit Margin
APATM	Adjusted Profit After Tax Margin
ROCE	Return on Capital Employed
RONW	Return on Net Worth

4.3 k-means Clustering Method

Nonhierarchical clustering techniques are designed to group *items*, rather than *variables*, into a collection of **k** clusters (Everitt and Dunn, 2001; Hair, Black, Babin and Anderson, 2010). The number of clusters, **k**, may either be specified in advance or determined as a part of the clustering procedure. The term k-means method is coined for describing an algorithm that assigns items to the **k**-clusters on the basis of nearest centroid (mean). Generally, this technique uses Euclidean distance measures between items computed by variables. Since the group labels are unknown for the data set, **k**-means clustering is one such technique in applied statistics that discovers acceptable classes i.e., groups of companies.

4.4 Multiple Discriminant Analysis

In this section, we focus on linear functions of financial ratios since Linear Discriminant Analysis (LDA) frequently achieves good results in the tasks of features extraction and object recognition, even though the assumptions of common covariance matrix among groups and normality are often violated (Duda, *et al.*, 2001). The initial idea of Linear Discriminant Analysis is to find linear transformation that best discriminate classes and classification is then performed in the transformed space based on metric such as Euclidean distance measure. Mathematically a typical Linear Discriminant Analysis is carried out through scatter matrix Analysis (Fukunaga, 1990). Fisher (1936) introduced LDA for two classes and its idea was to transform the multivariate observations X to univariate obserations Y such that the y 's derived from the two classes were separated as much as possible. Suppose that we have set of m, p - dimensional samples $X_1, X_2, ..., X_m$ (where $X_i = (x_{i1}, x_{i2}, ..., x_{ip})$) belonging to two different classes, namely c_1 and c_2. For the two classes the scatter matrices are given by

$$S_i = \sum_{x \notin c_i} \left(X - \overline{X_i}\right)\left(X - \overline{X_i}\right)'$$

4.5 Algorithm

A step-by-step algorithm for the performance of the industries (companies) during each of the study period based on their overall performances is described below:

Step 1: Factor analysis is initiated to find the structural pattern or data reduction underlying the data set.

Step 2: **k** –means analysis is used to partition the data set into **k**-clusters using the factor scores.

Step 3: Discriminant analysis is then performed with the original ratios by considering the groups formed by the **k**-means algorithm. The discriminant analysis is use to cross validate of k-means method and achieved their original classification.

5. RESULTS AND DISCUSSION

Factor analysis is performed first. As mentioned earlier, Varimax and Quartimax criterion for orthogonal rotation are used in factor analysis. Even though the results obtained by both the criterions are very similar, the varimax rotation provided relatively better clustering of financial ratios. Consequently, only the results of varimax rotation are reported here. We have decided to retain at least 70 percent of total variation in the data, and thus obtained consistently *four factors* for each year with eigen values little less than or equal to unity. *Table 2* shows variance accounted for by each factors.

Table 2. Percentage of variance explained by factors of different types of companies

Factor	Types of Companies				
	Cement	**Hardware and Software**	**Steel**	**Plastic**	**Leather**
1	39.468	36.389	37.148	38.357	38.357
2	13.890	17.439	14.914	14.117	14.117
3	11.221	10.197	13.778	12.971	12.971
4	8.546	7.853	8.985	9.446	9.446
Total	**73.12**	**71.87**	**74.82**	**74.89**	**77.40**

From the above table we observe that the total variances explained by the extracted factors are over 71 percent, which are relatively high. Also, for each factor the variability is more or less the same for the study period, though the number of industries for all the years, after data cleaning and selection, kept varying owing to various reasons. The financial ratios loaded in the factors are presented in *Table 3*. Only those ratios with higher loadings are indicated by bold numbers. From the *Table 3* it is clear that the factors of financial ratios are stable during the study period. We observe slight changes in factor loadings over the study periods considered. . The differences in factor loadings may be due to statistical variations in the original data.

Factor Analysis is generally used to find thriftiness among large number of variables. The quest is usually for underlying constructs or factors that explains each of the variation among several variables. In this study, fourteen financial ratios are used. After applying factor analysis it has reduced to four factors and are labeled as, *Cash Equity Ratio (PBITM, PBDTM, CPM, PBIDTM, RONW, ROCE, APATM), Liquidity Ratio (FIXED_AS, DEBTORS, CURRENT, LONG TERM),* *Profitability Ratio (INTEREST, INVENTOR) and Financial leverage Ratio (DEBT_EQU)* for each industries. Table 3 highlights the values, which are highly loaded in their respective four factors.

Table 3. Financial Ratios in Rotated Factors of different types of industries

Variable	Ratio Type	Factors : Cement industries			
		1	**2**	**3**	**4**
PBITM	Cash Equity Ratio	**.917**	-.076	-.200	.150
PBDTM		**.878**	-.064	-.263	.045
CPM		**.862**	-.279	.212	.021
PBIDTM		**.844**	-.287	-.226	.051
RONW		**.827**	.179	.365	-.112
ROCE		**.824**	.148	.388	-.135
APATM		**.697**	.386	.346	.198
INTEREST	Profitability	**.578**	.371	-.444	.382
INVENTOR		.077	**.727**	-.110	-.137
FIXED_AS	Liquidity	-.100	**.713**	.097	.056
DEBTORS		.395	**-.473**	.387	-.316
CURRENT	Liquidity	-.034	-.011	**.761**	.114
LONG_TER		-.050	.054	-.048	**.784**
DEBT_EQU	Financial Leverage Ratio	-.207	.161	-.322	**-.561**

Variable	Ratio Type	Factors : Hardware and Software			
		1	**2**	**3**	**4**
PBITM	Cash Equity Ratio	**.954**	-.085	-.073	-.087
PBDTM		**.948**	-.114	-.094	-.080
CPM		**.931**	-.112	.128	-.023
PBIDTM		**.828**	.357	.030	-.003
RONW		**.808**	.349	-.224	-.066
ROCE	Profitability	.431	**.811**	.077	-.093
APATM		.490	**.743**	-.025	-.137
INTEREST		-.199	**.728**	.085	.020
INVENTOR		.029	**-.484**	-.112	.030
FIXED_AS	Liquidity	.021	.032	**.820**	.052
DEBTORS		-.174	.372	**.816**	.036
CURRENT	Liquidity	.008	-.079	.060	**.859**
LONG_TER	Financial Leverage Ratio	.399	.263	-.141	**-.476**
DEBT_EQU		-.040	.365	-.355	**.418**

Variable	Ratio Type	Factors : Steel industries			
		1	**2**	**3**	**4**
PBITM	Cash Equity Ratio	**.967**	-.035	.089	.087
PBDTM		**.954**	-.017	.018	.093
CPM		**.896**	.142	-.055	-.104
PBIDTM		**.861**	.189	-.162	-.121
RONW		**.860**	-.132	.200	.164
ROCE	Profitability	**.544**	-.010	.230	.480
APATM		.059	**.981**	-.042	.015
INTEREST		.014	**.980**	.024	-.033
INVENTOR		-.161	-.077	**.671**	.047
FIXED_AS	Liquidity	-.262	.081	**.652**	.358
DEBTORS		.451	-.053	**.638**	-.188
CURRENT	Liquidity	.177	-.012	**.555**	-.173
LONG_TER	Financial Leverage Ratio	.532	.297	**.551**	-.065
DEBT_EQU		.054	-.025	-.148	**.884**

Variable	Ratio Type	Factors : Plastic industries			
		1	**2**	**3**	**4**
PBITM	Cash Equity Ratio	**.910**	.095	.060	.219
PBDTM		**.892**	.336	-.135	.161
CPM		**.888**	.252	-.137	.130

PBIDTM		**.691**	.592	.074	.023
RONW		.284	**.821**	-.004	.235
ROCE	Profitability	.221	**.806**	.045	.161
APATM		-.460	**.713**	-.061	.285
INTEREST		.481	**.711**	-.136	-.101
INVENTOR		.286	**.574**	-.146	-.016
FIXED_AS	Liquidity	-.067	-.049	**.985**	.004
DEBTORS		-.035	-.067	**.981**	-.048
CURRENT	Liquidity	-.100	-.058	-.041	**-.738**
LONG_TER	Financial Leverage Ratio	.042	-.004	-.086	**.673**
DEBT_EQU		.134	.223	.015	**.589**

Variable	**Ratio Type**	**Factors : Leather industries**			
		1	**2**	**3**	**4**
PBITM	Cash Equity Ratio	**.950**	.094	.139	.081
PBDTM		**.939**	-.033	.141	.097
CPM		**.933**	.101	-.115	.217
PBIDTM		**.923**	.122	-.089	.158
RONW		**.917**	-.060	.134	.168
ROCE		**.806**	.436	-.101	.075
APATM		**.733**	.604	-.022	.009
INTEREST	Profitability	.135	**.818**	.302	.131
INVENTOR		.000	**.618**	-.161	.237
FIXED_AS		.177	.063	**-.828**	.009
DEBTORS		.390	.202	**.588**	-.162
CURRENT	Liquidity	.215	.178	.094	**.734**
LONG_TER		.112	.185	-.237	**.683**
DEBT_EQU	Financial Leverage Ratio	.119	.005	.598	**.621**

After performing factor analysis, the next step is to assign initial group labels to each company. Step 2 of the algorithm is applied with factor scores extracted by Step 1, by conventional **k**-means clustering analysis. Formations of clusters are explored by considering 2-clusters, 3-clusters, 4-cluster and so on. Out of all the possible trials, 3-cluster exhibited meaningful interpretation than two, four and higher clusters. Having decided to consider only 3 clusters, it is possible to rate a company as Grade-**EP** (Elevated Performance), Grade-**MP** (Moderate Performance) or Grade-**SP** (Stumpy Performance) depending on whether the company belonged to Cluster 3, Cluster 2 or Cluster 1 respectively. Cluster 3 (Grade-**EP**) is a group of companies that have high values for the financial ratios, indicating that these companies are performing well.

The companies with lower values for the financial ratios are grouped into Cluster 1 (Grade-**SP**). This suggested that Cluster 1 is a group of companies with low-profile. Cluster 2 (Grade-**MP**) are those companies which perform moderately well. Even though analyses are done year-wise, only overall summary statistics are reported in *Table 4.*

Table 4. Number of companies in the clusters

Type of Industries	**Initial Cluster**			**Discriminant Classification**		
	1	**2**	**3**	**1**	**2**	**3**
Cement	003	270	507	003	422	355
Hardware and Software	208	379	253	174	377	289
Steel	386	162	602	383	157	610
Plastic	320	533	167	322	546	152
Leather	217	012	431	222	012	426

1 – Grade-**SP** 2 – Grade-**MP** 3 – Grade-**EP**

Table 4 indicates that Cement (99.6%) and Hardware & Software (79.2%) companies are mostly classified into MP and EP grades. Many of the companies of Steel (86.3%) and Leather (98.2%) category are placed into EP and SP grades. Companies associated with Plastic type (85.1%) are mostly classified into SP and MP grades. The possible reasons for such classification may be due to the then government policies. And also MNC's have found their way open business in India, pushing Indian companies back. *Figures 2* to *6* shows the groupings of industries into 3 clusters for each of the study period. It is interesting to note that the mean vectors of these clusters can be arranged in the order of magnitude.

In order to identify the factors that are mainly responsible for the formation of the groups, classification mapping is drawn using the standardized discriminant coefficients and the unstandardised discriminant functions evaluated at the group centroids. From the classification maps in *Figure 2 through 6*, it is evident that the three groups of rated companies are very well separated and represented for all the ten year periods of five types of industries.

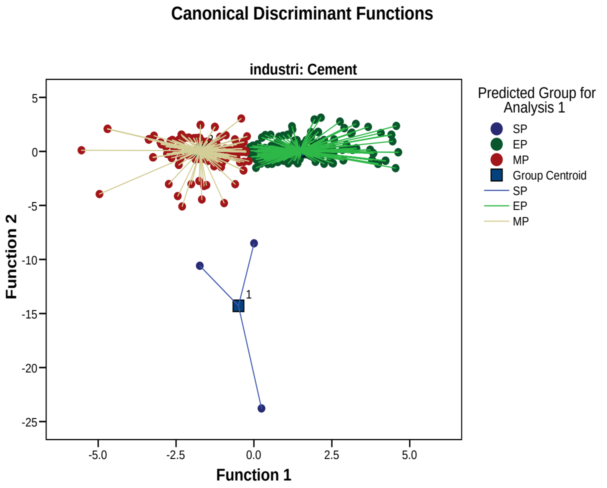

Figure 2. Three Groups of Cement Industries

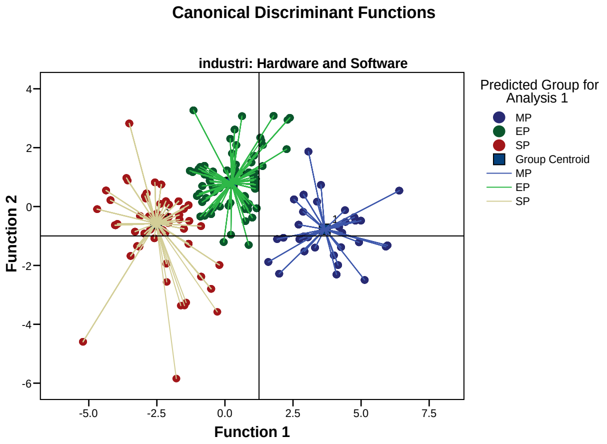

Figure 3. Three Groups of Hardware & Software Industries

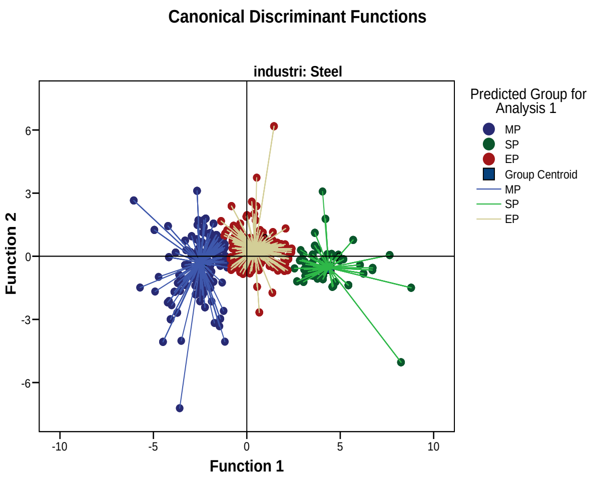

Figure 4. Three Groups of Steel Industries

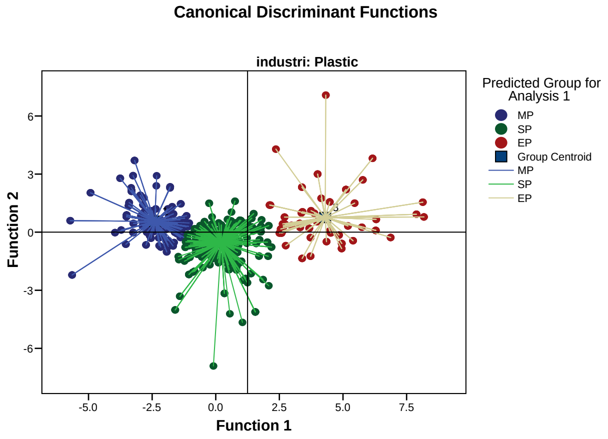

Figure 5. Three Groups of Plastic Industries

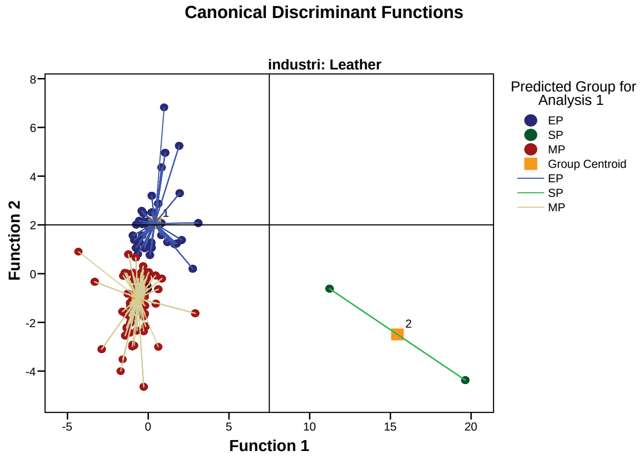

Figure 6. Three Groups of Leather Industries

6. CONCLUSION

The purpose of this paper is to identify the meaningful groups of industries that are classified as best with respect to their performance in terms of financial ratios and also to identify data reduction or hidden pattern of financial ratios using factor analysis, data mining and classification map techniques. An attempt is made to analyze the financial data relating to five major types of industries of public and private sector companies over a period of ten years from 2001 to 2010. The present analysis has shown that only 3 groups could be meaningfully formed for each type of industry. This indicates that only 3 types of companies existed over a period of ten years. Further, the industries find themselves classified into *Elevated* (Grade- **EP**), *Moderate* (Grade-**MP**) and *Stumpy* (Grade-**SP**) categories depending on the financial ratios. Factor analysis shows that fourteen financial ratios are reduced to four factors based on interrelation of variable and they are named as Cash equity Ratio, liquidity ratio, financial leverage ratio, profitability ratio. In each factor highly correlated variables accommodated with respective factor and some of the differences in factor loadings may be due to statistical variations in the original data. Financial Analyst can make use of these techniques of rating, and the industries can project the performance on the basis of financial ratios that has been considered in this study. A generalization of the results is under investigation to obtain a set of 3 groups of industries for any given year.

7. ACKNOWLEDGEMENT

We wish to thank Dr. G. Manimannan, Formerly Assistant Professor in the Department of Statistics, Madras Christian College, Tambaram, for fruitful discussions and support in analyzing the data.

8. REFERENCES

[1] Anderson T. W. (1984). An Introduction to Multivariate Statistical Analyis, 2/e, John Wiley and Sons, Inc., New York.

[2] Altman, E. I. (1968). Financial ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy, *Journal of Finance, 23(4), pp. 589-609*.

[3] Altman, E. I. (1994). An international Survey of Business Failure Classification Models, N.Y. University Salomon Center, *Volume 6, Number 2, 1997, p. 23.*

[4] Deakin, E. 1972. A discriminant analysis of predictors of business failure. Journal of Accounting Research 10(1): 167-179.

[5] M. Kantardzic, "Data Mining: Concepts, Models, Methods, and Algorithms", the textbook, IEEE Press & John Wiley, (First edition, November 2002; Second Edition, August 2011).

[6] Chandrasekaran R, Manimannan G and Lakshmi Priya (2014). Assessing Indian industries on the basis of Financial Ratios using certain Data Mining Tools, *International Journal of Statistika and Mathematika, India, Volume 9, Issue 2, 2014 pp 50-55.*

[7] Chandrasekaran R, Manimannan G and Lakshmi Priya (2013), International Journal of Scientific & Engineering Research Volume 4, Issue 2, ISSN 2229-5518.

[8] Chandrasekaran R, Manimannan G and Lakshmi Priya (2014). IOSR Journal of Mathematics (IOSR-JM), e-ISSN: 2278-5728, p-ISSN:2319-765X. Volume 10, Issue 1 Ver. IV, PP 63-71

[9] Everitt, B. and Dunn, G. (2001). Applied Multivariate Data Analysis, 2^{nd} Ed., Hodder Arnold, New York.

[10] Beaver, W. H. (1966). Financial ratios as Predictors of Failure, *Journal of Accounting Research, Vol 4. pp. 71-111*.

[11] Hair, J. F., Black, W. C., Babin, B. J., and Anderson, R. E. (2010). Multivariate Data Analysis, 7th ed.,Prentice Hall, New York.

[12] R.O. Duda, P.E. Hart, and D.G. Stork. Pattern Classification. John Wiley & Sons, New York, 2001.

[13] R.A. Fisher. The use of Multiple Measurements in Taxonomic Problems. Annals of Eugenics, 7:179–188, 1936.

[14] K. Fukunaga. Introduction to Statistical Pattern recognition. Academic Press, San Diego, CA, 1990.