Uludağ Üniversitesi
İktisadi ve İdari Bilimler Fakültesi Dergisi
Cilt XXIII, Sayı 2, 2004, s. 75-90

APPLYING KNOWLEDGE DISCOVERY AND DATA MINING TECHNIQUES ON DATA OF ERZURUM CHAMBER OF COMMERCE

Abdulkadir ÖZDEMİR*
Y. Ziya AYIK**

Abstract

Continuous growing data through out the world creates the problem of accessing the data in the required format and on the right time where it needed. Therefore the need for accessing and using the implicit knowledge in the data forces us to develop new software and hardware technologies. Consequently, Knowledge Discovery in Databases (KDD), one of the data processing techniques, has been appeared. Data Mining (DM) has been developed as a method acquiring knowledge from the data in knowledge discovery.

In this study, database and building data warehouse are explained, and theoretical information is given for the KDD. As an application study, first the steps of the knowledge discovery in database are applied on a data warehouse which is built by using real data of members of Erzurum Chamber of Commerce, second some hidden information in the data is discovered and finally the results are organized as a report.

Keywords: *Data, Knowledge, Database, Data Warehouse, Knowledge Discovery in Databases (KDD), Data Mining (DM).*

---

* *Uzm. Dr.; Atatürk Üniversitesi Mühendislik Fakültesi, Elektrik ve Elektronik Müh. Böl., Erzurum. akadir@atauni.edu.tr*

** *Yrd. Doç. Dr.; Atatürk Üniversitesi Erzurum Meslek Yüksek Okulu, Bilgisayar Teknolojileri ve Programlama Bölümü, Erzurum. ziyaayik@atauni.edu.tr*

Özet

Bilgi Keşfi ve Veri Madenciliği Yöntemlerinin Erzurum Ticaret Odası Verileri Üzerine Uygulanması

Dünyada verinin böyle hızlı bir şekilde artması, verilere gerektiği yerde, gerektiği zaman ve gerektiği şekilde erişebilme sorununu ortaya çıkarmaktadır. Bu nedenle veriye erişim ve veride saklı bilginin kullanımı yeni yazılım ve donanım teknolojilerinin geliştirilmesini zorunlu kılmıştır. Bunun sonucu olarak veri işleme tekniklerden biri olan Veritabanlarından Bilgi Keşfi (VTBK) ortaya çıkmıştır. Veri madenciliği (VM), veritabanlarından bilgi keşfinde, veriden bilgi elde etme yöntemi olarak geliştirilmiştir.

Yapılan çalışmada veritabanı ve veri ambarının oluşturulması konuları açıklanmakta, VTBK için gerekli olan teorik bilgiler verilmekte, bu kavramların veritabanı ve veri ambarı ile ilişkileri ortaya koyulmaktadır. Uygulama çalışmasında ise Erzurum Ticaret Odası'ndan alınan oda üyelerine ait gerçek verilerden, veri ambarı oluşturularak, veritabanlarından bilgi keşfi adımları uygulanmakta, veriler içinde keşfedilmemiş anlamlı bilgiler ortaya çıkarılmakta ve bu bilgiler son olarak rapor şeklinde düzenlenmektedir.

Anahtar Kelimeler: *Veri, Bilgi, Veritabanı, Veri ambarı, Veritabanlarında Bilgi Keşfi (VTBK), Veri Madenciliği (VM).*

INTRODUCTION

Knowledge and data production are growing exponentially because of improving computer and related technologies. Therefore, our time in which knowledge and data are produced and shared continuously is called information era. What is important for the people of information era is not the large amount of the knowledge they have, but their ability to use improved technologies to reach the required information.

In global world where competition is increasing, detailed research is gaining importance to improve the profit and productivity. What is the most needed for these research studies is in fact hidden in the data that the companies have. Hence, the companies that have activities in different fields have chosen to save all their data in database form. Today's computer technologies should solve the problem of not only saving the huge amount of data in the most economical way but also picking up the significant knowledge in it. There are problems in the detection of the significant knowledge among huge data bulk because of the limited time and constraints on the sources. In order to overcome these problems, techniques to obtain knowledge and information out of data have been developed.

Nowadays one of the most remarkable techniques is "Knowledge Discovery in Databases (KDD)". Although "Data Mining (DM)" is a single step of this technique, it is sometimes used to express the whole technique. KDD is in fact a collective method which is a combination of different methods. During coding of these techniques, different algorithms such as neural networks and genetic algorithms have been used. Therefore, these algorithms require the use of state of art computer resources.

1. DEFINITIONS

Although concepts such as data, knowledge and information are well known by everyone, sometimes they are used instead of each other by mistake. To clear this confusion it's important to define frequently used concepts in knowledge discovery and DM briefly.

Data, Knowledge and Information: All of the stored components are called data, things obtained from data are called knowledge and things obtained for a decision support are called information **(**Akpınar, 1997: 7).

Database: The data produced by devices or people which is stored in an organized manner is called database (Uysal, 2000: 7), (Riccardi, 2001: 16).

Data warehouse: The database, including mostly combined or summarized data related to the past and structured for query, is called data warehouse (Lehtenbörger and Vossen, 2003: 416). In other words, data warehouse deviates from the database due to the fact that it is stable and includes the data regarding the past. Creating a new database using the data coming from the snapshots of an existing database in different times is named as building a data warehouse. On the other hand, a data warehouse which is small and structured for a special purpose is called data mart (Adriaans and Zantinge, 1998: 25).

Pattern: Rules, not previously known, stated in simple/ comprehensive manner during the DM are called pattern **(**Witten and Frank, 2000: 3).

2. KNOWLEDGE DISCOVERY AND DATA MINING TECHNIQUES

Today, it is a common fact that it is getting hard to follow the information since more data, than one can read in his whole life, can be produced weekly or hundreds of megabyte of data can be distributed over the internet daily (Chen etc. 1996: 866). Exponentially increasing amount of the data introduces some difficulties when it comes to acquire the required data within the whole. This is like looking for a needle in 'a big haystack',

except that the hay stack is expanding by time (Adriaans and Zantinge, 1998: 2). This leads to the fact that the KDD should be able to follow the data increase rate (Zhang, 1999: 475).

When the KDD is not performed, it is possible that either hidden information can not be extracted or a duplicate/wrong record can not be identified properly. This may cause the problem of wasting of the available resources (Westphal and Blaxton, 1998: 7).

It is suggested that KDD can be performed in six steps (Adriaans and Zantinge, 1998: 37), (Berson etc., 1999: 203), (Olaru and Wehenkel, 1999: 20), (Fayyad etc., 1996: 10). These steps are as follows;

- Data selection and organization,
- Data cleaning,
- Data enrichment,
- Data coding,
- DM,
- Reporting.

2.1. Data Selection and Organization

First the data of the KDD application for the data warehouse should be selected and then the following adjustments should be made on the data (Westphal and Blaxton, 1998: 81).

- The text fields in the data should be converted to either upper case or lower case.
- Fields, saved separately but belonging to the same record, should be combined in a field.
- Possible different data formats should be converted to a format which is suitable for the purpose.
- Unqualified characters/entries should be cleaned.
- Some fields which have text format should be coded in a numerical form.
- Different units should be converted to a unique unit for the purpose.

2.2. Data Cleaning

In this step it is aimed to decrease the data pollution level and to complete the missing data. In the DM, in order to get satisfactory results one of the best options is possibly to get rid of the missing data which is

impossible to be completed. Erroneous data may exist because of spelling error, typo or different spelling options. These types of data should be corrected. If this corrupts the data, they should be removed from the data set (Westphal and Blaxton, 1998: 101).

2.3. Data Enrichment

This step is performed to get a single table which has all required entries for the DM by using alternate accessible data or databases (Westphal and Blaxton, 1998: 89).

2.4. Data Coding

Sometimes using data with their real values may reveal a pattern creation problem. If, for example, the data have continuous values, then a set of predefined range of values should be coded (Westphal and Blaxton, 1998: 36).

2.5. Data Mining

Since the beginning of 1990s, researchers in this area have defined the DM as a process of extracting the information that exists in the data but undiscovered and perhaps useful (Fayyad, 1996: 21), by means of unruled procedures (Famili etc, 1997: 4). The DM is not a correlation, statistics or sorting process applied on a data set. It is rather a process of creating a new pattern or model by means of unevaluated data which can not be obtained by traditional analysis. It is also defined as a technique that requires human perception, patience and flexibility on repetitive processes (Westphal and Blaxton, 1998: 16).

The DM is a process which requires repetitive procedures based on the inspection of its suitability of the obtained results to the targeted goal. Thus, it is not a single phase study (Ryu and Eick, 2005: 30). In some cases, the DM may need to be repeated by adjusting the system set-up and expanding or narrowing the data set (Alpaydın, 2003: 10).

In the DM, all techniques that help to extract more knowledge from the data are useful. Although different techniques are available for various aims, the followings are the most common used ones (Adriaans and Zantinge, 1998: 47).

- Query Tools,
- Statistical Techniques,
- Visualization Techniques,
- On Line Analytical Processing (OLAP),

- Case-Based Learning,
- Decision Trees,
- Association Rules,
- Neural Networks,
- Genetic Algorithms (Hu, 2005: 1).

On the other hand, some DM analysis, where these techniques are used, can be listed as following (Fu, 1997: 18), (Akpınar, 2000: 5).

Association analysis: In this type of analysis, negative and positive correlations are examined between the data.

Regression analysis: The goal is to use the results, obtained from the current conditions of the data, in later processes.

Classification analysis: This analysis is similar to the regression analysis. However, class values are used when the data values are continuous.

Clustering analysis: Here, the clusters, made up by the data, are determined, and the features belonging to these clusters are defined.

Sequential pattern analysis: Predictions for the future can be obtained from the existing data distributed in time.

Fraud detection analysis: In this analysis, the data, that show important differences from the most common data are examined. This analysis facilitates catching the frauds.

2.6. Reporting

Reporting the obtained results is the most important step for the evaluation and use of the study as it is the case in all studies. However, reporting the results, belonging to a database that DM is applied, is perhaps the most difficult step of the KDD. Information to be reported should be suitable for the perception levels and expectations of the people that demand knowledge. Thus, when preparing the final reports, questions such as “who is the target group for the report?”, “what are their interests?”, “what part of the report are they interested in?”, “what is their expertise?” should be taken into account (Westphal and Blaxton, 1998: 12). The report prepared by considering these questions will draw more interest and attention.

3. APPLYING KNOWLEDGE DISCOVERY AND DATA MINING TECHNIQUES ON THE DATA OF ERZURUM CHAMBER OF COMMERCE

Erzurum Chamber of Commerce (ECC), with approximately 5000 registered members, has a database where the data are entered by using

FileMaker. Some of the member records are illustrated in Table-1. Due to legal considerations, private data such as name of the company and income tax are taken out of the database. Data mart is constructed by taking into account ethical rules and the laws. Since the amount of data is large enough that can be processed by a personal computer a PC, with Pentium 4 and 2.8 GHz processor is used. In addition, Microsoft Excel 2003 software is anticipated as data mart. The process for applying knowledge discovery steps on ECC data is as follows.

3.1. Data Selection and Ordering

First, required data are chosen by removing the data perceived as unnecessary for several reasons. The data are organized so that different knowledge in the same data field is separated into two fields. For instance, data field that has the combined information for the name of province and member ID is separated into two fields as name of the province (1) and member ID (2). Following Excel commands are used for this separation process.

=if(isnumber(A2);"MERKEZ";mid(A2;1;(find("/";A2)-1))) (1)

=if(isnumber(A2);A2;value(mid(A2;(find("/";A2)+1);5))) (2)

Second, since the field that includes record of the capital has both alphanumeric and numeric formats, first alphanumeric data are converted into numeric data then all amounts in this field are converted into YTL (New Turkish Lira).

Last, the different formats of typing are eliminated in the data field. For instance, expressions such as “ANONİM ŞİRKET”, “ANONİM ŞİR-KETİ”, “A.Ş.” are formatted into “ANONİM ŞİRKET”.

3.2. Data Cleaning

Companies that have no capital record, branch offices and the records kept for only the member of cooperatives with very low capitals such as only 1 or 2 YTL are removed from data ware at this step. As last step, fields such as title address and phone number of the company are removed.

3.3. Data Enrichment

Erzurum and its provinces’ population records, taken from Governmental Statistical Institute (DİE) (http://www.die.gov.tr/nufus_sayimi/2000tablo5.xls), are added to their

related fields in the data mart. New profession names, listed in Table-2, are created by taking into account company type and its working field.

3.4. Data Coding

Since capital data have a vast range, capital data are coded by using intervals as shown in Table-3 in order to prevent getting misleading results.

3.5. Data Mining

There are about 60 different DM software (Özdemir, 2004: 87). Most of these are target-specific and application-based software. Some of them can be used over the Internet for research purposes. Following are some of the DM software that has been offered as a demo or trial version; Clementine, Darwin, DataCruncher (DataMind), DBMiner, Enterprise Miner, GainSmarts, Intelligent Miner, MineSet (SGI), Model-1 and NeuroShell (Elder and Abbott, 1998: 8).

DataCruncher software of Data Mind Firm (http://www.datamindcorp.com), which has all features except recording capability of the outputs, and one-month trial version of MineSet software of Silicon Graphics Inc. (http://www.sgi.com/Products/software/MineSet/), distributed over the internet, are used in this study. Classification and clustering analysis can be done with the DataCruncher software (Figure-1). On the other hand, the MineSet software has useful features of visual mining applications (Figure-2).

The finalized version of the data prepared for the DM is tabulated in Table-4. Since the data of ECC was not too much in quantity and they were not collected in accordance with the data warehouse format, some analysis performed over these data did not give satisfactory results. Some of the analyses with the satisfactory results are as follows:

1. Clustering analysis of Erzurum and its provinces at financial level,
2. Clustering analysis of capital at company level,
3. Correlation of jobs with population at province level,
4. Variation of capital with provinces.

3.6. Reporting

1) Clustering Analysis of Erzurum and Its Provinces at Financial Level: Clustering analysis is performed with the DataCruncher software. The corresponding results, shown in Figure 3, are as follows:

The software grouped the capital data in five clusters. According to this, Erzurum companies fall into mid and upper level whereas its province companies fall into mid and lower level capital companies.

2) Clustering Analysis of Capital at Company Level: This analysis is also performed with the DataCruncher software. The corresponding results, shown in Figure 4, are as follows:

Company data shows five clusters. When compared "Kollektif" (corporate), "Hakiki Şahıs" (person), "Anonim" (incorporated) and "Limitet" (limited) companies mostly fall into mid and lower level, uniformly distributed, lower level and upper level capital companies, respectively.

3) Correlation of Jobs with Population at Province Level: It is performed with the MineSet software which enables to work with more cluster groups (Figure-5). The corresponding results are as follows:

In Erzurum downtown there are all types of business companies, but with the decreasing population the variety of the companies is decreasing in the provinces. For example, the wholesale food, engineering contractor and gas station companies exist almost in all downtowns whereas mining companies exist only in Erzurum downtown.

4) Variation of Capital with Provinces: The last analysis, performed with the MineSet is shown in Figure 6. The results obtained are:

Capital structure is organized in an irregular manner according to the provinces and company types. Textile companies are more common in Horasan, İspir and Tortum, and are owned by a small capital. The biggest companies are wholesale food, engineering reseller and gas station companies.

CONCLUSION

Advances in computer and communication techniques enable these technologies to be used in all areas. Switching from hardcopy filing to computerized filing or switching from mailing and hardcopy correspondence to network based communication technologies are good examples of these advances. Also, outstanding growth of raw data is another factor for the improvement of new techniques in data processing applications. Today, one of the techniques, improved for that purpose, is DM. It enables us to extract the knowledge, which can not be obtained with the traditional methods, out of big amount of data. Some problems that may exist in knowledge extracting from big amount of data lead to the improvement of KDD, which is a combination of a group of techniques. In other words, this method is

composed of the DM, data warehouse and a series of processes, and it simplifies the process of extracting knowledge out of data in big quantities.

In practice, the data of ECC, which is increasing in volume, is used, and the DM is applied to this data. It is shown that in this organization, where big amount of data is processed, important and useful knowledge can be obtained. Therefore, it is important to establish a data warehouse. In such a data warehouse, classification and clustering analysis are performed; company structure, activity field, capital, region and distribution according to population for companies in Erzurum and its provinces' are investigated; and the obtained results are summarized in a report in accordance with the KDD technique. Some records or fields of the data has to be excluded in the study because of data secrecy, ethic rules and some collected data inconsistent with data warehouse structure. More informative evaluation will be possible in the future by recording these data in data warehouse format and getting more flexible data usage permissions in the ECC database.

REFERENCES

Adriaans, P. and Zantinge, D. (1998), **Data Mining**, Addison Wesley Longman, England.

Akpınar, H. (1997) "Enformasyon Teknolojisi ve İşletmecilik Öğretimine Etkileri", http://www.isletme.istanbul.edu.tr/akpinar/content /Enformasyon%20Teknolojileri.pdf, İstanbul Üniversitesi İşletme Fakültesi, 2003, s.:1-45.

Akpınar, H. (2000) "Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği", **İstanbul Üniversitesi, İşletme Fakültesi Dergisi**, vol:29/Nisan , 1-22.

Alpaydın, E. (2003) "Zeki Veri Madenciliği" (Notes), Bilişim 2000 Eğitim Semineri Notları, http://www.cmpe.boun.edu.tr/~ethem/files/paper_/veri-maden_2k-notlar.ppt, Boğaziçi Üniversitesi.

Berson, A., Smith, S. and Thearling, K. (1999), **Building Data Mining Applications for CRM**, McGraw-Hill, USA.

Chen, M., Han, J. and Yu, P. S. (1996), "Data Mining: An Overview from a Database Perspective", **IEEE Transactions on Knowledge and Data Engineering**, Vol. 8 No. 6, December, 866-883.

Elder, J. F. and Abbott, D. W. (1998), "A Comparison of Leading Data Mining Tools", Elder Research, http://www.datamininglab.com/pubs/kdd98_elder_abbott_nopics_bw.pdf, New York, USA.

Famili, A., Shen, W., Weber, R. and Simoudis, E. (1997), "Data Preprocessing and Intelligent Data Analysis", **Elsevier Intelligent Data Analysis 1**, October, 3-23.

Fayyad, U. M. (1996), "Data Mining and Knowledge Discovery: Making Sense Out of Data", **IEEE Expert**, October, 20-25.

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P. and Uthurusamy, R. (1996), **Advances in Knowledge Discovery and Data Mining**, AAAI Press / The MIT Press, USA.

Fu, Y. (1997), "Data Mining - Task, Techniques and Applications-", **IEEE Potentials**, October/November, 18-20.

Hu, Y. (2005), "Finding useful fuzzy concepts for pattern classification using genetic algorithm", **Elsevier Information Sciences** 175, 1-19.

Lehtenbörger, J. and Vossen, G. (2003), "Multidimensional normal forms for data warehouse design", **Elsevier Science Information Systems** 28, 415-434.

Olaru, C. and Wehenkel, L. (1999), "Data Mining", **IEEE Computer Applications in Power**, July, 19-25.

Özdemir, A. (2004), Veritabanlarında Bilgi Keşfi ve Veri Madenciliği, Atatürk Üniversitesi Sosyal Bilimler Enstitüsü, **Unpublished PhD Thesis**, Erzurum.

Riccardi, G. (2001), **Principles of Database Systems with Internet and Java Applications**, Addison-Wesley Publishing Company, USA.

Ryu, T. and Eick, C. (2005), "A database clustering methodology and tool", **Elsevier Science Information Sciences** 171, 29-59.

Uysal, M. (2000), SQL **Veri Tabanı Sorgulama Dili**, Beta Basım Yayım Dağıtım A.Ş., 3. Edition, İstanbul.

Westphal, C. and Blaxton, T. (1998), **Data Mining Solutions: Methods and Tools for Solving Real-World Problems**, Wiley Computer Publishing, USA.

Witten, I., H. and Frank, E. (2000) **Data Mining: Practical Machine Learning Tools and Techniques with JAVA Implementations**, ABD, Morgan Kaufmann Publishers.

Zhang, S. (1999) "Aggregation and Maintenance for Database Mining", **Elsevier Intelligent Data Analysis,** V. 1, 475-490.

TABLES

Table 1. Commerce Chamber Database Example

…	⋮	⋮	⋮	⋮	⋮
Capital	⋮	50,000.-TL	20,000.-TL	30,000.-TL	⋮
Registering Date	⋮	08.03.1948	04.04.1950	12.08.1959	⋮
Company Type	⋮	HAKİKİ ŞAHIS	HAKİKİ ŞAHIS	HAKİKİ ŞAHIS	⋮
Address	⋮	TAŞMAĞAZALAR CAD…	TEBRİZKAPI…	HINIS; CUMHURİYET CAD.	⋮
Birth Date	⋮	1923	1922	1928	⋮
Birth Place	⋮	TORTUM	ERZURUM	HINIS	⋮
Title	⋮	N. G.	S. T.	S. G.	⋮
Chamber ID	⋮	2799	2967	5304	⋮
Member ID	⋮	374	459	HINIS/25	⋮

Table 2. Occupation Group Codes

Occupation Grp.	Occupation Group Name
1	AKARYAKIT-LPG
2	HAYVAN VE URUNLERI
3	UN VE UNLU MAMULLER
4	MANIFATURA, GIYIM
5	MOBILYA,DOGRAMA
6	KIRTASIYE, SAGLIK
7	BANKA, KOOPERATIF
8	NAKLIYAT, INSAAT
9	TOPTAN GIDA
10	MUTEAHITLIK
11	SOSYAL TESISLER
12	MADENCILIK
13	TICARET-TUCCAR
14	INSAAT MALZEMELERI
15	OTOMOBIL VE TICARETI
16	DAYANIKLI TUKETIM
17	ZIRAI KOOPERATIF
18	ELEKTRIK TAAHHUT

Table 3. New Turkish Lira (YTL) Coding Table

YTL Ranges	Code
0,00 - 1,00	1
1,01 - 10,00	2
10,01 - 100,00	3
100,01 - 1.000,00	4
1.000,01 - 10.000,00	5
10.000,01 - 100.000,00	6
100.000,01 - 1.000.000,00	7
1.000.000,01 - 10.000.000,00	8
10.000.000,01 - 100.000.000,00	9

Table 4. Prepared Data for Data Mining

Province	Member ID	Chamber ID	Company Type	Capital	Capital Range (YTL)	CapitalGrp	JobGrp Name	Population
…	…	…	…	…	…	…	…	…
ERZURUM	8447	13503	HAKIKI SAHIS	5.000.000	10,00	3	MOBILYA, DOGRAMA	361.235
NARMAN	73	11377	KOLL.STI.	600.000	1,00	2	AKARYAKIT-LPG	9.025
ERZURUM	12915	18610	HAKIKI SAHIS	50.000.000.000	100.000,00	7	AKARYAKIT-LPG	361.235
HINIS	230	18053	HAKIKI SAHIS	400.000.000	1.000,00	5	AKARYAKIT-LPG	27.504
TEKMAN	114	17835	HAKIKI SAHIS	400.000.000	1.000,00	5	AKARYAKIT-LPG	6.350
…	…	…	…	…	…	…	…	…

FIGURES

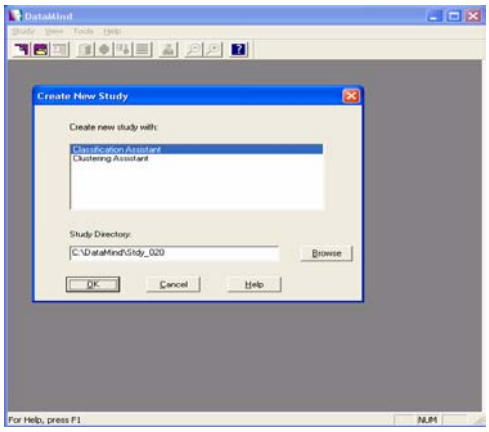

Figure 1. *DataCruncher Software*

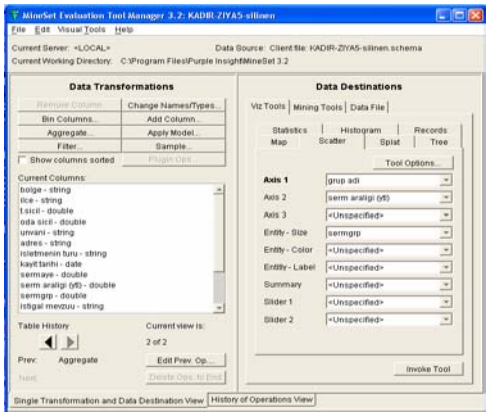

Figure 2. *MineSet Software*

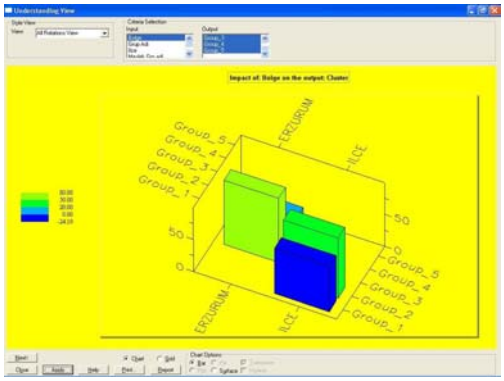

Figure 3. *Clustering analysis of Erzurum and its provinces at financial level*

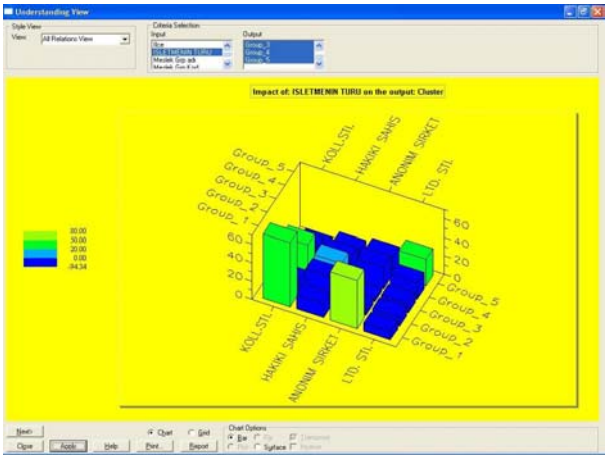

Figure 4. *Clustering analysis of capital at company level*

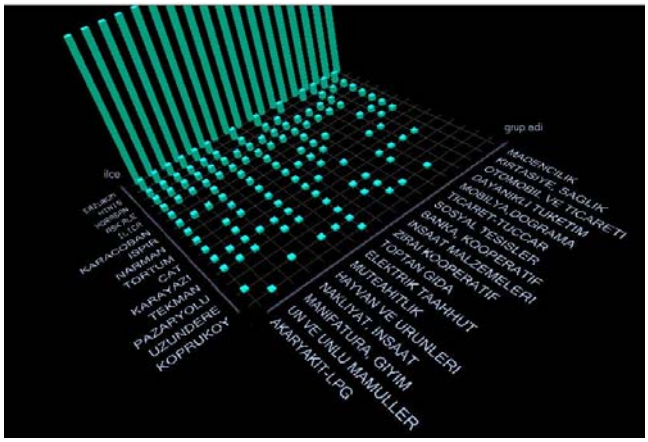

Figure 5. *Correlation of jobs with population at province level*

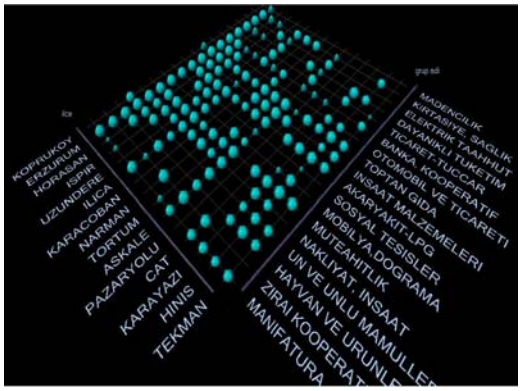

Figure 6. *Variation of capital with provinces*