The Research for the Factors Affecting on the Success of the Students by Using Apriori and Decision Tree Algorithms

U. Ekim, G. Tezel, and H. Kodaz

***Abstract*—In recent years, datum have reached great dimensions and obtaining significant and useful information from these datum have come to be a very difficult operation that takes a long time by human's power. In the completion of this operation with an easy and a fast manner, it has been seen that the concept of data mining carries a big importance. In this Study, making a survey for the preparatory class students of Selcuk University by the web, it has been tried to find the factors affecting on the success of the students by the data mining methods. Therefore, from the methods of data mining, it has been used apriori and decision tree algorithms. In the Study, a survey has been made for the students by the web, and the questions of this survey done and the answers given for these questions have been held in a digital environment. As for these answers, after analysing by two different programmes prepared by in MATLAB media, some laws have been found, and the results have been compared.**

***Index Terms*—Apriori, Decision Tree, Data Mining.**

I. Introduction

By the datum held in computer environment recently show increase in a large amount, it has been emerged the need of benefitting from these datum efficiently. Therefore, data mining consisting of the stages of setting and evaluating of the model has gained a big importance. From the inside of datum in a large amount, the searching of the rules, which will provide us to estimate about future, by the help of computer programmes is called as "data mining". New generation hardware and software have emerged from the deficiency in interpreting raw datum in a large amount. The knowledge exploration in the database is to develop new generation instruments and techniques that interpret datum in a large dimension with a half or complete automatic manner. Data mining, in the process of knowledge exploration in the databases, consists of the stages of setting and evaluating of the model. Data mining used for many objectives from the classification of datum to making a decision has started to gain more importance since nineties (Alpaydın, 2000; Jackson, 2002).

In this Study, it has been mentioned about the association rules from data mining techniques. In the research, from the association rules, it has been used apriori and decision tree algorithms; and these algorithms have been prepared on MATLAB. Here, it has been tried to find the factors affecting on the success of the preparatory class students at Selcuk University during their educational life.

Manuscript received June 25, 2012; revised September 13, 2012.

The authors are with the Department of Computer Engineering, University of Selcuk, Konya, Turkey (e-mail: hkodaz@selcuk.edu.tr, hkodaz@selcuk.edu.tr).

II. Material and Method

Before the application, a prestudy has been done. Firstly; a survey has been applied on the students that are known their success standings, by the web. In this survey, there are seven questions. The questions and answers of the survey have been numerically held in the database. In this study, the results have been analyzed and compared after both algorithms were applied on the answers which were randomized as 200 persons from among the students, gave. In this way, it has been aimed at measuring the effect of differences in the number of student on the results required by the study. With apriori and decision tree algorithms obtained with MATLAB, analyzing the answers that the students gave, it has been sought to find the factors affecting on the success of successful students.

A. Apriori Algorithm

In apriori algorithm in which factor sets are frequently used with searching the database, it is reached frequently used factor sets that provide one-element minimum support metric in the first search. In the later searches, the frequently used factor sets founded in the previous search are used for producing new potential frequently used factor sets that are called as "candidate sets" (Agrawal et al. 1993, Agrawal and Srikant 1994).

During the search, it is calculated the support metric of the candidate sets. The sets frequently used and providing minimum support metric are taken the candidate sets off. The frequently used factor sets become a candidate set for a next passing. Until there is not any frequently used factor set, this process continues like that. If k-factor set provides for minimum support metric, as the main approach in the apriori algorithm, the subsets of this set provide support metric, too (Sever and Oğuz, 2002). Numerical product codes are used in the market basket data as is ascending sort. If the factor sets called together with element numbers have k piece product, it is showed with k-factor set. In order to provide support metric for every factor set for the factor sets of which product codes are ascending sort, a counter variable has been attached. When a factor set is constituted at the first time, counter variable is zeroized (Agrawal et al., 1993; Agrawal and Srikant, 1994).

B. Decision Tree Algorithm

Decision tree algorithm is a practical method used widely

for an effective deduction. Decision trees provide that the examples are lined in an order like a tree from root to leaf (Quinlan 1986; Mitchell, 1997; Kshertapalapuram and M. Kirley, 2005; Dehuri and Cho, 2008).

1) ID3 algorithm

ID3 algorithm, which benefits from the concept of "entropy", is to be able to find the variable that has the most distinctive feature in the classification among other variables. The concept of "entropy" known as the digitization of the present information is used for measuring the indefiniteness and the randomness inside the data set. Entropy taking a value between 0–1, when all possibilities are equal, reaches its biggest value (Quinlan 1986; Mitchell, 1997; Silahtaroglu, 2008).

Entropy, mathematically is defined in the following way. p-variable, when it indicates the possibilities from 1 to n (Equation 1)

$$H(p_1, p_2, .., p_n) = \sum_{i=1}^{n} p_i log(1/p_i) \quad (1)$$

It's like this.

First of all; taking the whole of positive and negative datum registered in the database into account, the entropy of the whole of database is calculated. If the database is also divided into sub-divisions, then it is calculated the entropy of these sub-divisions one by one. After reached the entropy of the database, it is reached the root part and the leaves of the tree structure. The value of entropy that is reached for the whole of database and the values reached for every different variable inside the datum are found separately. Every result found is called as "acquisition".

$$Retrieval(D,S) = H(D) - \sum_{i=1}^{n} P(Di)H(Di) \quad (2)$$

The big one among the acquisitions found is selected as the root part of the tree. Once again, according to the same equation (Equation 2.), the leaves of the tree are found. In this way, the tree structure is constituted (Quinlan 1986; Mitchell, 1997).

III. The Mechanism of the Application

The hundred fold of the ratio of the number of student calculated to find that the answers coexist against the total number of student gives the coexistence ratio of the answers. The calculation of this ratio will be like in Equation 3.

$$CRA = \frac{NSC}{TNS} * 100 \quad (3)$$

CRA: Coexistence Rate of Answers

NSC: The number of student counted to find this ratio

TNS: The Total Number of Students

In the apriori algorithm applied, when minimum support for a group having 200 persons is taken as 2 and 3, the results appeared are such as in Table I and Fig. 1.

TABLE I: The Survey Group Having 200 Persons and the Rules Obtained for Minimum Support Value 2.

Questions Coexisting	Status	Min. Sup. Value 2		Min. Sup. Value 3	
		NSC	CRA	NSC	CRA
Father; educator and mother; housewife	successful	99	53,8	60	53,5
	unsuccessful	85	46,1	52	46,4
Mother; housewife, father graduate student	successful	98	57,3	93	55,3
	unsuccessful	73	42,6	75	44,6
Father a high school graduate, mother primary school graduate	successful	99	49,5	62	39,7
	unsuccessful	101	50,5	94	60,2
Mother; graduate student and family's monthly income is 2000 TL	successful	99	54,3	64	55,1
	unsuccessful	83	45,6	52	44,8
Family monthly income is above 2000 TL and science high school graduate student	successful	96	56,8	33	51,5
	unsuccessful	73	43,1	31	48,4
General	successful	491	54,1	312	50,6
	unsuccessful	415	45,8	304	49,3

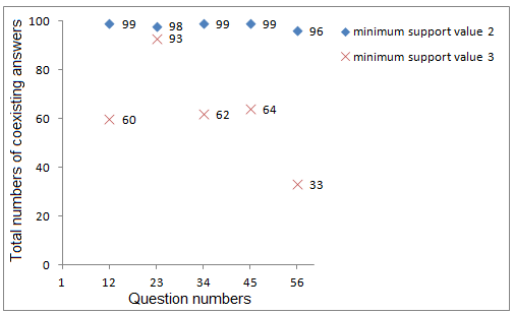

Fig. 1. The survey group having 200 persons and figure of the rules obtained for minimum support value 2 and minimum support value 3.

The results that are gotten by ID3 algorithm from the decision tree algorithms is applied on a group having 200 persons are as in Table II. The general results obtained from both algorithms have been indicated in Table III.

In general, when the results emerged from both algorithms are compared with each other, in the answers coexisting, it is seen that the results emerged from apriori algorithm for successful students according to the coexistence ratio of the answers are higher than the results emerged from the decision tree, once again.

TABLE II: RULES OBTAINED WHEN ID3 ALGORITHM IS APPLIED ON THE GROUP HAVING 200 PERSONS.

Questions Coexisting	Status	NSC	CRA
Father; educator and mother; housewife	successful	72	52,5
	unsuccessful	65	47,4
Mother; housewife, father graduate student	successful	68	56,1
	unsuccessful	53	43,8
Father a high school graduate, mother primary school graduate	successful	47	43,1
	unsuccessful	62	56,8
Mother; graduate student and family monthly income is above 2000 TL	successful	52	53,6
	unsuccessful	45	46,3
General	successful	239	51,5
	unsuccessful	225	48,4

TABLE III: GENERAL RESULTS FROM APRIORI AND DECISION TREE ALGORITHMS.

Groups	successful	unsuccessful	Generally Coexisting ratio of Answers %
200 persons groups, minimum support = 2	491		54,1
		415	45,8
200 persons groups, minimum support = 3	312		50,6
		304	49,3
When decision tree algorithm is applied on the Group having 200 persons	239		51,5
		225	48,4

IV. CONCLUSION

For apriori algorithm application, in a group having 200 persons, it is seen that the coexistence ratio, generally, is smaller than the half of one's participating in the survey. It is previously known whether the students taking part in the survey are successful or not. Here, the survey results show that how low the coexistence ratios of the answers emerged after they are evaluated by algorithms, namely, the coexistence ratios of the questions are. In other word, the numbers of students giving the same answers are more than the questions and in this way, the general ratio is high are called "coexistence".

The general coexistence success rate of a group having 200 persons has been seen as higher in comparison to another group. The general coexistence ratio of a group having 200 persons according to minimum support 2 and 3 has come in view as near to each other; as for another group, it has been seen that minimum support 2 is better than minimum support 3. In respect of decision tree algorithm, it has appeared that the general coexistence ratio of a group having 200 persons according to another group has been higher; however, in general, as is in the apriori algorithm, the general coexistence ratio of both groups are lower than its half.

At the results emerged from the program done with both algorithms, it is seen that the answers of the coexisting questions, according to their coexistence rates, are lower than the half of the general coexistence ratio. However, it appears that the coexistence ratio of the results emerged from the apriori algorithm are higher in comparison to the results emerged from the decision tree algorithm. In both of the results too, in general, the questions of "what is mother's profession", "...father and mother's education level" and "...'s high school graduated" come into prominence. Here, it is seen how students' family lives before university affect on their successes in the future. Particularly, it is understood that mothers' professions are a big factor for children's successes in the future. When looked at the results emerged by taken the minimum support as 2 in a group having 200 persons, it appears that 98 students whose mother is a housewife and whose father is a graduate student have a coexistence ratio like % 57,3. Again, when looked at the results emerged by taken the minimum support as 3 in this group, it is seen that 93 students whose mother is a housewife and as for whose father is a graduate student have a coexistence ratio as % 55,3. When ID3 algorithm from the decision tree algorithms is applied on a group having 200 persons, it is seen that 68 students whose mother is a housewife and whose father is a graduate student reach for the highest coexistence ratio in its group with a % 56,1 coexistence. As is in the group having 200 persons, 91 students whose mother is a housewife and whose father is a graduate student reach for the highest coexistence ratio in this group with a % 53,5 coexistence.

Consequently, in this study, it has been tried to reach the factors affecting the successes of students in the preparatory class by apriori and decision tree algorithms from the data mining algorithms. It has been used the questions and the answers of the survey study that are applied on the students in the preparatory class with this objective. After the questions and the answers given of this survey were

transformed into numerical values, apriori and decision tree methods have been applied by MATLAB program; and according to the intensity of the answers given, it has been researched which factors positively or negatively affect on students during their educational life.

ACKNOWLEDGEMENT

This study has been supported by Scientific Research Project of Selcuk University.

REFERENCES

[1] E. Alpaydın, "Zeki veri madenciliği: ham veriden altın bilgiye ulaşma yöntemleri," *Bilişim2000 eğitim semineri*, 2000.

[2] S. Dehuri and S. Cho, "Multi-objective Classification Rule Mining Using Gene Expression Programming," in *Proc. of Third 2008 International Conference on Convergence and Hybrid Information Technology*. Washington, USA. 2008, vol. 7, pp. 54-760.

[3] K. K. Kshertapalapuram and M. Kirley, "Mining Classification Rules Using Evolutionary Multi-objective Algorithms," in *Proc. of Knowledge-Based Intelligent Information and Engineering Systems*, Part 3, Springer. 2005, pp. 959- 965.

[4] J. Jackson, "Data Mining: A Conceptual Overview," *Communications of the Association for Information Systems*, 2002, vol. 8, pp. 267-296.

[5] R. Agrawal and T. Imielinski, "A. Swami. Mining Association Rules between Sets of Items in Large Databases," in *Proc. of the 1993 ACM SIGMOD International Conference on Management of Data*, 1993.

[6] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules," in *Proc. 20th Int. Conf. Very Large Data Bases, VLDB*. 1994.

[7] T. M. Mitchell, "Machine Learning," C. L. Liu, B. T. Allen, McGraw-Hill Companies, Inc., Carnegie Mellon University, Pensilvanya. 1997, pp. 52-53.

[8] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, no. 1, pp. 81–106, 1986.

[9] H. Sever and B. Oğuz, "Veri Tabanlarında Bilgi Keşfine Formel Bir Yaklaşım," *Bilgi Dünyası*, vol. 3, no. 2, pp. 173-204, 2002.

[10] G. Silahtaroğlu, "Kavram ve Algoritmalarıyla Temel Veri Madenciliği," *Papatya Yayıncılık*, Istanbul, 2008.