A Comparison of Statistical Dependency and Functional Dependency between Attributes Based on Data

HYONTAI SUG
Department of Computer Engineering,
Dongseo University,
47 Jurye-ro, Sasang-gu, Busan, 47011,
REPUBLIC OF KOREA

Abstract: - Chi-squared test is a standard statistical test to ascertain independence between categorical variables. So, it is recommended to do the test for the attributes in the datasets, and remove any redundant attributes before we supply the datasets to machine learning algorithms. But, if we have many attributes that are common in real-world datasets, it is not easy to choose two attributes to do the independence test. On the other hand, several automated algorithms to find functional dependencies based on data have been suggested. Because functional dependencies show many-to-one relationships between values of attributes, we could conjecture that there might be statistical dependence in the found functional dependencies. For us to overcome the problem of choosing appropriate attributes for statistical dependency tests, we may use some algorithms for automated functional dependency finding. We want to confirm that the found functional dependencies can show statistical dependence between attributes in real-world datasets. Experiments were performed for three different real-world datasets using SPSS to confirm the statistical dependence of functional dependencies that are found by an open-source tool called FDtool, where we can use FDtool for automated functional dependency discovery. The experiments confirmed that there exists statistical dependence in the found functional dependencies and showed improvements in decision trees after removing dependent attributes.

Key-Words: - Artificial intelligence, machine learning, classification, statistical independence, functional dependency, knowledge modeling, preprocessing, relations, data tables

Received: May 23, 2021. Revised: August 29, 2022. Accepted: September 24, 2022. Published: October 25, 2022.

1 Introduction

Determining functional dependency is an important theoretical basis for the normalization of relational databases. If we look at the purpose of normalization, it is to store one fact in one place in the databases, thereby minimizing potential errors by storing data redundantly. The functional dependencies in the attributes of relations are determined by checking whether they have a many-to-one relationship with all the values that may appear in the attributes. And such a decision is made by the database designer, [1]. On the other hand, there is a series of research works to find out such functional dependencies automatically from stored data. Finding functional dependencies based on stored data may require a lot of computing time unless we apply some elaborate algorithms because we can have exponential combinations of attributes to consider. Automated algorithms try to find functional dependencies as efficiently as possible, so efficient algorithms have been suggested, [2]. But, functional dependencies found by data may not be real functional dependencies. Let's see an example.

We have a shipment relation like table 1. S# and P# mean supplier number and part number respectively.

Table 1. A shipment relation

S#	P#	Quantity
S1	P1	100
S1	P2	100
S2	P1	200
S2	P2	200

Then, we have functional dependencies based on data like;

{S#, P#} -> {Quantity},
{S#} -> {Quantity},
{Quantity} -> {S#}.

But, we know that the first one only is the real functional dependency, while the others are not.

On the other hand, a statistical method to check dependency between categorical attributes is the chi-squared test, [3]. It is recommended that when we have a contingency table of 2 by 2, we should have all expected frequencies > 5, and for a larger table, no more than 20% of all cells may have an

expected frequency < 5, and all expected frequencies > 1, [4]. For example, assume that we surveyed replies for a policy by 100 men and women. We can have a contingency table like table 2.

Table 2. A Contingency table of a survey

	yes	no	total
men	28	16	44
women	21	35	56
total	49	51	100

We can determine whether the replies are dependent on sex with the chi-squared test. Because of their good understandability, decision trees are important data mining tools when human wants to understand the constructed knowledge models like the bio-medical domain, and this kind of dependency check between conditional attributes is important for classification task of data mining like decision trees that are considered one of the most important machine learning algorithms, [5]. The target datasets of decision trees consist of conditional attributes and decisional attributes. The precondition of conditional attributes is that they are independent of each other and dependent on decisional attributes only. Because decision trees have the property of low bias but high variance, [6], [7], which means that decision trees reflect the composition of data themselves very well, if we have some dependency between conditional attributes, we may have more complex trees or the final knowledge models, [8], [9]. So, J.R. Quinlan recommended getting rid of dependency between conditional attributes before we begin to generate decision trees by doing the dependency check of the chi-squared test before in his book of C4.5, [10]. But, if we have many attributes in conditional attributes, we can have exponential combinations of attributes, so statistical dependency checking can be very time-consuming.

Because found functional dependencies based on data represent the relationship of many to one between attributes, there could be regularities in the occurrence of values of the attributes. In other words, there could be statistical dependence between the attributes. Therefore, we want to find out the statistical significance of functional dependencies in real-world datasets by doing the chi-squared test for the same attribute sets which are found to have functional dependencies. As a result, we may save some time for the statistical dependency check, otherwise, if we have tried exhaustively on the time-consuming tasks.

2 Related Work

Functional dependencies are used to determine the many-to-one relationship of values between the attributes of relations or relation variables, [11], and serve as a criterion for judgment for normalization. That is, a relation variable will be in the second normal form if it is in the first normal form and every non-key attribute is fully functionally dependent on the primary key. And a relation variable will be in the third normal form if it is in the second normal form and it does not contain any transitive functional dependency. It is recommended that a relation variable should be at least in the third form in a practical sense. Since database designers sometimes make mistakes in database design, several algorithms have been proposed to discover functional dependencies using data stored in relations or data tables. We can refer to the found information for further normalization. The developed algorithms try to find functional dependencies as efficiently as possible so several efficient algorithms have been suggested, [2], [12], [13]. On the other hand, unlike traditional functional dependencies, functional dependencies limited to specific values of attributes, called conditional functional dependencies, are used for data cleaning purposes, [14], [15]. Moreover, approximate conditional functional dependencies (ACFD) which can be applied to the subsets of tuples in relation have been suggested, [16]. The format of functional dependencies of ACFD has similarity with class association rules, [17], [18].

Because functional dependencies represent many-to-one correspondences of attribute values including one-to-one relationships of values that appear in a relation, if there is a functional dependency, there may be some statistical dependency between the two attributes or between the two sets of attributes that appear in the left-hand side (LHS) and the right -hand side (RHS) of the functional dependency. The chi-squared test is a well-known method for statistical independence for categorical attributes. So, we may want to do statistical tests for found functional dependencies based on data.

We need to discretize continuous or numerical attributes before we apply the chi-squared test because we can apply the test for categorical attributes only. Many discretization algorithms have been suggested, and García et al. compared 80 different discretization algorithms, [19]. For example, a simple method for discretization is called binning. Binning is the process of dividing a continuous variable into a constant interval and converting each interval into a value for a new

categorical variable. Among the many discretization methods, we will use a method based on information entropy and minimum description length (MDL) principle by Fayyad and Irani, [20]. Their discretization method finds the best split in which the bins are as pure as possible. The purity of a bin is measured by the ratio of the values in a bin having the same class label. The method is characterized by finding the split with the MDL that is based on entropy. Because we want to see the effect of eliminating redundant attributes with a well-known decision tree algorithm, C4.5, that is also based on the entropy and MDL principle. C4.5 is known as one of the top 10 data mining algorithms, [21].

3 Problem Formulation

To do an independent test that determines that two categorical variables are statistically related or independent of each other, chi-squared test is often used, [22]. In statistics, we usually use the terminology, categorical variables, while in computer science we use the terminology, nominal attributes, so the two terminologies have the same meaning.

Let's see a simple example that needs the independence test. We have a categorical variable called educational level with 5 categorical values, such as primary school graduate, secondary school graduate, high school graduate, college degree, and graduate degree, and a categorical variable called annual income with 3 values of the upper, middle, and lower class. You can use a chi-squared test for a problem such as determining whether they are independent of each other or not. We can create a contingency table and do a chi-squared test to see the relation.

If the two variables that classify data are called X and Y, and the variables X have m and the variable Y has n categories or different values, then an m×n contingency table can be created. In the contingency table, the element at row i and column j called, O_{ij}, represents the observations corresponding to the i_{th} category of X, and the j_{th} category of Y.

Since the null hypothesis H_0 that two variables are independent is known to approximate the chi-squared distribution with the degree of freedom (m-1)(n-1), if the observations O_{11}, O_{21}, ..., and O_{mn} do not differ from the corresponding expected frequency E_{11}, E_{21}, ..., and E_{mn}, the value of the test statistic will be 0. Conversely, if the difference between the degree of observation and the expected frequency is large, the value of the test statistic will also be large, so H_0 will not be established. That is, we can make a statistical judgment that the two variables are not independent. The definition of functional dependency based on data can be defined as follows, [11].

Definition 1. Let **r** be a relation over the set of attributes U, and A, B be any subset of U. Then B is functionally dependent on A, A → B, if and only if each A value in **r** is associated with precisely one B value. □

Based on the above definition, we can make a contingency table for the relation **r** as follows:
If the two sets of attributes A and B over **r**, where the attribute set A has m and the attribute set B has n different values, then an m×n contingency table can be created. In the contingency table, O_{ij} represents the observations corresponding to the i_{th} value of A, and the j_{th} value of B. □

In the case of a chi-squared test, there is the inconvenience of having to specify two attributes to perform the test. On the other hand, because the algorithm to find functional dependencies by data automatically can find all the functional dependencies, we may use the automated algorithm complementarily with the chi-squared test to select the two attributes. However, when the size of the data is small, relatively much more functional dependencies will be found compared to large-sized data, and some of them can be fake functional dependencies because of the property of functional dependency of many to one relationship as we can see from the simple example relation in table 1. Therefore, large datasets are preferred for more real functional dependencies based on data. Moreover, fewer columned datasets are not for our interests because our final goal is to find simpler decision trees of which the complexity of the trees is affected largely by the many conditional attributes.

M. Buranosky et al. implemented an automated program called FDtool that finds functional dependencies in the datasets of tabular form, [23]. FDTool is a Python-based open-source program to mine functional dependencies and candidate keys. The number of attributes of the input data is limited to 26. For the statistical test, we will use a well-known tool, IBM SPSS, [24], for our experiment. For discretization and decision tree training we will use an open-source tool called Weka, [25].

3.1 Experimental Procedure

We want to check whether a given dataset consisting of conditional and decisional attributes for data mining has functional dependencies and statistical dependencies between the conditional attributes. To check functional dependencies, we use FDtool, then, if some functional dependencies are

found, we try to do chi-squared tests for the set of attributes in the found functional dependencies, before we supply the dataset to generate decision trees. The detailed procedure for our experiment is as follows:

EXPERIMENTAL PROCEDURE:
INPUT: a dataset D in tabular form.
BEGIN

1. Discretize continuous attributes if they are contained in conditional attributes in D;
2. Find functional dependencies based on data in conditional attributes using FDtool;
3. **For** each found functional dependency **Do**
   Do the chi-squared test between LHS and RHS of found functional dependency unless RHS has only one value;
   End For;
4. **For** each LHS and RHS of found functional dependency that is decided to have statistical dependency **Do**
   Let S be LHS or RHS of the found functional dependency;
   Remove the columns of attributes in S from D;
   Generate a decision tree;
   End For;
5. Select the best decision tree regarding the size and accuracy from the result of 4.

END.

In the procedure, LHS and RHS mean the left-hand side and the right-hand side of the found functional dependencies respectively. Because FDtool can find functional dependencies based on data, we may have functional dependencies of RHS consisting of only one value, that is, a constant, and in that case, we cannot do chi-squared tests as indicated 3rd in the procedure.

4 Experimentation

Because FDtool limits the number of input attributes, and we want datasets having a mix of continuous and nominal attributes as well as having a lot of instances, there are not many public datasets that meet our criteria. So, we chose three public datasets, called adult, bank, and credit approval datasets from the UCI machine learning repository for our experiments, [26]. The datasets consist of 14 ~ 16 conditional attributes and one decisional attribute. Because the chi-squared test can be done between nominal attributes, we need to discretize any continuous attributes in the datasets. We apply FDtool after discretization to find functional dependencies between conditional attributes, and we expect many functional dependencies based on stored data in each dataset.

4.1 Adult Dataset

The purpose of the adult dataset is a census dataset to predict whether income exceeds $50K/year depending on various aspects. The data set has 48,842 records and has 14 conditional attributes and one decisional attribute, named 'class' having two different values, >50K or <=50K. The 14 conditional attributes consist of numerical and categorical attributes as in table 3. Categorical attributes have nominal values, while numerical attributes have numbers, so we need to discretize them. We applied Fayyad and Irani's multi-valued discretization method implemented in Weka. Table 3 shows the results. In the notation '(' means '<', and ']' means '<='. For example, (21.5~23.5] means '21.5 < an interval <= 23.5'.

Table 3. Discretized conditional attributes of the adult dataset

ATTRIBUTE	VALUES
age	Numeric => (-∞~21.5], (21.5~23.5], (23.5~24.5], (24.5~27.5], (27.5~30.5], (30.5~35.5], (35.5~41.5], (41.5~54.5], (54.5~61.5], (61.5~67.5], (67.5~ ∞]
workclass	8 nominal values (Private, …, Never-worked)
fnlwgt	Numeric => 'all'
education	16 nominal values (Bachelors, … , Preschool)
education-num	Numeric => (-∞~8.5], (8.5~9.5], (9.5~10.5], (10.5~12.5], (12.5~13.5], (13.5~14.5], (14.5~ ∞]
marital-status	7 nominal values (Married-civ-spouse, … , Married-AF-spouse)
occupation	14 nominal values (Tech-support, …, Armed-Forces)
relationship	6 nominal values (Wife, …, Unmarried)
race	5 nominal values (White, …, Black)
sex	2 nominal values (Female, Male)
capital-gain	Numeric => (-∞~57], (57~3048], (3048~3120], (3120~4243.5], (4243.5~4401], (4401~4668.5], (4668.5~4826], (4826~432.5], (4932.5~4973.5], (4973.5~5119], (5119~5316.5], (5316.5~5505.5], (5505.5~5638.5], (5638.5~6389], (6389~6457.5], (6457.5~6505.5], (6505.5~6667.5], (6667.5~70555.5], (7055.5~ ∞]
capital-loss	Numeric => (-∞~1551.5], (1557.5~1568.5], (1568.5~1820.5],

	(1820.5~1834.5], (1834.5~1846], (1846~1859], (1859~1881.5], (1881.5~1894.5], (1894.5~1927.5], (1927.5~1975.5], (1975.5~1978.5], (1978.5~2168.5], (2168.5~2176.5], (2176.5~2218.5], (2218.5~2310.5], (2310.5~2364.5], (2364.5~2384.5], (2384.5~2450.5], (2450.5~2469.5], (2469.5~3089.5], (3089.5~ ∞]
hours-per-week	Numeric (-∞~34.5], (34.5~39.5], (39.5~41.5], (41.5~49.5], (49.5~61.5], (61.5~ ∞]
native-country	41 nominal values (United-States, …, Holand-Netherlands)

4.1.1 Checking Functional Dependencies for Adult Dataset

14 functional dependencies in the conditional attributes were found as follows:

{age} -> { fnlwgt}
{ workclass} -> { fnlwgt}
{education} -> {education-num}
{education} -> { fnlwgt}
{education-num} -> { fnlwgt}
{marital-status} -> { fnlwgt}
{occupation} -> { fnlwgt}
{relationship} -> { fnlwgt}
{race} -> { fnlwgt}
{sex} -> { fnlwgt}
{capital-gain} -> { fnlwgt}
{capital-loss} -> { fnlwgt}
{hours-per-week} -> { fnlwgt}
{native-country} -> { fnlwgt}

Because functional dependency between attributes represents many to one relationship in attributes values, the found functional dependencies show such relationships, but attribute fnlwgt has only one value, 'all' as we can see from the third row in table 3, the functional dependencies that have fnlwgt in left-hand side (LHS) or right-hand side (RHS) are meaningless for the chi-squared test. Therefore, only the functional dependency, {education} -> {education-num} is left to do the statistical test. Note that the attribute education has 16 different values, while the attribute education-num has 7 different values as in table 3. Chi-squared test was done for the two attributes, education, and education-num. The contingency table (cross table) of the two attributes can be summarized in table 4. Note that the original cross table is 16×7, but only one column has no zero value in each row of the table. So, for notational convenience and easy understanding, it is summarized as shown in table 4. We can see that the values of attribute education have many to one relationship to attribute education-num as shown in table 4.

Table 4. Corresponding values in the cross table of the two attributes, education and education-num in the adult dataset

education	education-num	frequency
10th	(-∞~8.5]	1389
11th	(-∞~8.5]	1812
12th	(-∞~8.5]	657
1st-4th	(-∞~8.5]	247
5th-6th	(-∞~8.5]	509
7th-8th	(-∞~8.5]	955
9th	(-∞~8.5]	756
Assoc-acdm	(10.5~12.5]	1601
Assoc-voc	(10.5~12.5]	2061
Bachelors	(12.5~13.5]	8025
Doctorate	(14.5~ ∞]	594
HS-grad	(8.5~9.5]	15784
Masters	(13.5~14.5]	2657
Pre-school	(-∞~8.5]	83
Prof-school	(14.5~ ∞]	834
Some-college	(9.5~10.5]	10878
TOTAL		48842

The following table 5 shows the result of the chi-squared test of the two attributes, education and education-num. Because the value of asymptotic significance is 0.0 and Pearson chi-square is very large, we can decide the two attributes are dependent.

Table 5. The result of the chi-squared test of the two attributes, education and education-num in the adult dataset

	Value	Degree of freedom	Asymptotic significance(2-sided)
Pearson Chi-square	293052.0	90	0.0

4.1.2 Generating Decision Trees

Table 6 shows the property of the decision tree generated by J4.8 which is a version of C4.5 written in Java in Weka from the original and discretized adult dataset. All experiments are performed in 10-fold cross-validation.

Table 6. Decision tree from the discretized adult dataset

Number of leaves	522		
Size of the tree	584		
Accuracy	86.743%		
		Predicted >50K	Predicted ≤50K

Confusion matrix	Actual >50K	7226	4461
	Actual ≤50K	2014	35141

We try to generate decision trees for the dataset having select attributes only. Table 7 shows the property of the decision tree generated from the adult dataset of which attribute education-num which is RHS of the functional dependency {education} -> {education-num} is omitted.

Table 7. Decision tree from the discretized adult dataset of which education-num attribute is omitted

Number of leaves	577		
Size of the tree	644		
Accuracy	86.6447%		
Confusion matrix		Predicted >50K	Predicted ≤50K
	Actual >50K	7143	4544
	Actual ≤50K	1976	35179

Table 8 shows the property of the decision tree generated from the adult dataset of which attribute education which is LHS of the functional dependency {education} -> {education-num} is omitted.

Table 8. Decision tree from the discretized adult dataset of which education attribute is omitted

Number of leaves	516		
Size of the tree	578		
Accuracy	86.7515%		
Confusion matrix		Predicted >50K	Predicted ≤50K
	Actual >50K	7218	4469
	Actual ≤50K	2002	35153

As we compare table 7 and table 8, omitting attribute education has some better effect in reducing the size of the tree as well as some increase of the accuracy. Note that the type of values of the attribute education consists of more variety than the attribute education-num. In other words, we have many-to-one relationships in the functional dependency, {education} -> {education-num}, and there is a variety of them as summarized in table 9. This is the reason why we have a better decision tree when we omit attribute education.

Table 9. Attribute values of education and education_num that make up the many-to-one relationship

education	education-num	frequency
10^{th}, 11^{th}, 12^{th}, 1^{st}-4^{th}, 5^{th}-6^{th}, 7^{th}-8^{th}, 9^{th}, Pre-school	(-∞~8.5]	6408
Assoc-acdm, Assoc-voc	(10.5~12.5]	3602
Bachelors	(12.5~13.5]	8025
Doctorate, Prof-school	(14.5~ ∞]	594
HS-grad	(8.5~9.5]	15784
Masters	(13.5~14.5]	2657
Some-college	(9.5~10.5]	10878
TOTAL		48842

4.2 Bank Dataset

The purpose of the bank dataset is to predict if the client will subscribe to a term deposit for direct marketing campaigns of a Portuguese banking institution. The data set has 45,211 records and 16 conditional attributes and one decisional attribute, named ‘y’, having two different values, yes or no. The 16 conditional attributes have a variety of values as in table 10, where the values of numeric attributes are discretized.

Table 10. Discretized conditional attributes of bank dataset

ATTRIBUTE	VALUES
age	Numeric => (-∞~25.5], (25.5~29.5], (29.5~60.5], (60.5~ ∞]
job	12 nominal values (admin., …, services)
marital	3 nominal values (divorced, married, single)
education	4 nominal values (Primary, …, unknown)
default	2 nominal values (no, yes)
balance	Numeric => (-∞~46.5], (46.5~105.5], (105.5~1578.5], (1578.5~ ∞]
housing (housing loan)	2 nominal values (no, yes)
loan (personal loan)	2 nominal values (no, yes)
contact	3 nominal values (cellular, …, unknown)
day	Numeric => (-∞~1.5], (1.5~4.5], (4.5~9.5], (9.5~10.5], (10.5~16.5], (16.5~21.5], (21.5~25.5], (25.5~27.5], (27.5~29.5],

	(29.5~30.5], (30.5~ ∞]
month(last contact month of year)	12 nominal values (jan, ..., dec)
Duration(last contact duration, in seconds)	Numeric => (-∞~77.5], (77.5~130.5], (130.5~206.5], (206.5~259.5], (259.5~410.5], (410.5~521.5], (521.5~647.5], (647.5~827.5], (827.5~ ∞]
Campaign(number of contacts performed during this campaign)	Numeric => (-∞~1.5], (1.5~3.5], (3.5~11.5], (11.5~ ∞]
Pdays(number of days that passed by after the client was last contacted)	Numeric => (-∞~8.5], (8.5~86.5], (86.5~99.5], (99.5~107.5], (107.5~177.5], (177.5~184.5], (184.5~203.5], (203.5~316.5], (316.5~373.5], (373.5~ ∞]
Previous(number of contacts performed before this campaign)	Numeric => (-∞~0.5], (0.5~ ∞]
Poutcome(outcome of the previous marketing campaign)	4 nominal values (failure, …, success)

4.2.1 Checking Functional Dependencies for Bank Dataset

One functional dependency was found in the conditional attributes based on the dataset as follows: {poutcome, pdays} -> {previous}

Based on the found functional dependency between conditional attributes, we can do chi-squared tests between the attribute sets, {poutcome, pdays} and {previous}. Before we give input for SPSS, we combined the values of the two attributes, the poutcome, and the pdays, named pdays_poutcome, row by row. Chi-squared test was done for the attributes. The cross table of the attributes can be summarized in table 11. Even though pdays has 10 nominal values and poutcome has 4 nominal values, there are no rows corresponding to the values, (177.5~184.5]unknown, (203.5~316.5]unknown, (316.5~373.5]unknown, (8.5~86.5]unknown, and (99.5~107.5]unkown. As a result, table 11 has 35 rows only. Table 11 shows the many-to-one relationship between attribute pdays_poutcome and attribute previous except the second row which represents the one-to-one correspondence of values. We can see that the summary of the cross table of the bank dataset has a very simple many-to-one relationship of values compared to those of the adult dataset in table 4.

Table 11. Corresponding values in the cross table of the two attributes, pdays_poutcome and previous in bank dataset

pdays_poutcome	previous	frequency
(-∞~8.5]failure	(0.5~ ∞]	10
(-∞~8.5]unknown	(-∞~0.5]	36954
(-∞~8.5]other	(0.5~ ∞]	83
(-∞~8.5]success	(0.5~ ∞]	15
(107.5~177.5]failure	(0.5~ ∞]	971
(107.5~177.5]unknown	(0.5~ ∞]	1
(107.5~177.5]other	(0.5~ ∞]	318
(107.5~177.5]success	(0.5~ ∞]	124
(177.5~184.5]failure	(0.5~ ∞]	275
(177.5~184.5]other	(0.5~ ∞]	68
(177.5~184.5]success	(0.5~ ∞]	287
(184.5~203.5]failure	(0.5~ ∞]	346
(184.5~203.5]unknown	(0.5~ ∞]	1
(184.5~203.5]other	(0.5~ ∞]	131
(184.5~203.5]success	(0.5~ ∞]	173
(203.5~316.5]failure	(0.5~ ∞]	1116
(203.5~316.5]other	(0.5~ ∞]	402
(203.5~316.5]success	(0.5~ ∞]	123
(316.5~373.5]failure	(0.5~ ∞]	1377
(316.5~373.5]other	(0.5~ ∞]	506
(316.5~373.5]success	(0.5~ ∞]	66
(373.5~∞]failure	(0.5~ ∞]	185
(373.5~∞]unknown	(0.5~ ∞]	2
(373.5~∞]other	(0.5~ ∞]	69
(373.5~∞]success	(0.5~ ∞]	67
(8.5~86.5]failure	(0.5~ ∞]	224
(8.5~86.5]other	(0.5~ ∞]	92
(8.5~86.5]success	(0.5~ ∞]	138
(86.5~99.5]failure	(0.5~ ∞]	285
(86.5.5~99.5]unknown	(0.5~ ∞]	1
(86.5~99.5]other	(0.5~ ∞]	139
(86.5~99.5]success	(0.5~ ∞]	420
(99.5~107.5]failure	(0.5~ ∞]	112
(99.5~107.5]other	(0.5~ ∞]	32
(99.5~107.5]success	(0.5~ ∞]	98
TOTAL		45211

The following table 12 shows the result of the chi-squared test of the two attributes, pdays_poutcome and previous. Because the value of asymptotic significance is 0.0 and Pearson chi-square is very large, we can decide the two attributes are dependent.

Table 12. The result of the chi-squared test of the two attributes, pdays_poutcome and previous in the bank dataset

	Value	Degree of freedom	Asymptotic significance(2-sided)
Pearson Chi-square	45211.0	34	0.0

4.2.2 Generating Decision Trees

Table 13 shows the property of the decision tree from the original, but discretized bank dataset. All experiments are performed with 10-fold cross-validation.

Table 13. Decision tree from discretized bank dataset

Number of leaves	671		
Size of the tree	807		
Accuracy	90.3032%		
Confusion matrix		Predicted yes	Predicted no
	Actual yes	2253	3036
	Actual no	1348	38574

We'll try to remove attributes from the original dataset in the following order; the pdays and the poutcome together, and the previous based on the found functional dependency, {poutcome, pdays} -> {previous}. Table 14 shows the property of the decision tree generated from the bank dataset of which the attributes poutcome and pdays are omitted.

Table 14. Decision tree from bank dataset minus pdays and poutcome attribute

Number of leaves	835		
Size of the tree	1033		
Accuracy	89.8653%		
Confusion matrix		Predicted yes	Predicted no
	Actual yes	2212	3077
	Actual no	1505	38417

Table 15 shows the property of the decision tree generated from the bank dataset of which the attribute previous is omitted.

Table15. Decision tree from bank dataset minus previous attribute

Number of leaves	671		
Size of the tree	807		
Accuracy	90.3032%		
Confusion matrix		Predicted yes	Predicted no
	Actual yes	2253	3036
	Actual no	1348	38574

If we compare table 13 which is from the original data and table 15 above, we have the same result so that removing a redundant attribute has no effect in reducing the size or improving the accuracy of the decision tree. This is because functional dependency is a very simple many-to-one relationship.

4.3 Credit approval Dataset

The purpose of the credit approval dataset is to decide positive or negative decisions for credit card applications. The data set has 690 records and 15 conditional attributes and one decisional attribute, named 'class', having two different values, + or -. The dataset has been provided in the form of all attribute names and values being changed to meaningless symbols to protect confidentiality. The 15 conditional attributes have a variety of values as in table 16, where the values of numeric attributes are discretized.

Table 16. Discretized conditional attributes of credit approval dataset

ATTRIBUTE	VALUES
A1	2 nominal values (a, b)
A2	Numerical => (-∞~38.96], (38.96~ ∞]
A3	Numerical => (-∞~4.2075], (4.2075~ ∞]
A4	4 nominal values (u, y, l, t)
A5	3 nominal values (g, p, gg)
A6	14 nominal values (c, d, cc, i, j, k, m, r, q, w, x, e, aa, ff)
A7	9 nominal values (v, h, bb, j, n, z, dd, ff, o)
A8	Numerical => (-∞~1.02], (1.02~ ∞]
A9	2 nominal values (t, f)
A10	2 nominal values (t, f)
A11	Numerical => (-∞~0.5], (0.5~2.5], (2.5~ ∞]
A12	2 nominal values (t, f)
A13	32 nominal values (g, p, s)
A14	Numerical => (-∞~105], (105~ ∞]
A15	Numeric => (-∞~492], (492~ ∞]

4.3.1 Checking Functional Dependencies for Credit Approval Dataset

Three functional dependencies were found in the conditional attributes based on the dataset as follows:

{A4} -> {A5}
{A5} -> {A4}
{A11} -> {A10}

Based on the found functional dependencies between conditional attributes, we can do chi-squared tests between the attribute sets, {A4, A5} and {A10, A11}. The cross table of the attributes {A4, A5} can be summarized in table 17. When we read table 17, we may be confused if there are one-to-many relations between A5 and A4. But, because we have found two functional dependencies, A4 -> A5 as well as A5 -> A4, they must be one-to-one. In the table '?' value means unknown value. Even though A4 and A5 have 4 and 5 different values, the dataset has only four combinations of values as in table 17.

Table 17. Corresponding values in the cross table of the two attributes, A4 and A5 in the credit approval dataset

A4	A5	frequency
?	?	6
l	g	2
u	g	519
y	p	163
TOTAL		690

The following table 18 shows the result of the chi-squared test of the two attributes, A4 and A5. Because the value of asymptotic significance is 0.0 and Pearson chi-square is very large, we can decide the two attributes are dependent.

Table 18. The result of the chi-squared test of the two attributes, A4 and A5 in the credit approval dataset

	Value	Degree of freedom	Asymptotic significance(2-sided)
Pearson Chi-square	1380.0	6	0.0

The cross table of the attributes {A10, A11} can be summarized in table 19. Even though A10 and A11 have 2 and 3 different values, the dataset has only three combinations of values as in table 19.

Table 19. Corresponding values in the cross table of the two attributes, A10 and A11 in the credit approval dataset

A11	A10	frequency
(-∞~0.5]	f	395
(0.5~2.5]	t	116
(2.5~ ∞]	t	179
TOTAL		690

The following table 20 shows the result of the chi-squared test of the two attributes, A10 and A11. Because the value of asymptotic significance is 0.0 and Pearson chi-square is very large, we can decide the two attributes are dependent.

Table 20. The result of the chi-squared test of the two attributes, A10 and A11 in the credit approval dataset

	Value	Degree of freedom	Asymptotic significance(2-sided)
Pearson Chi-square	690.0	2	0.0

4.3.2 Generating Decision Tree

Table 21 shows the property of the decision tree from the original, but discretized credit approval dataset. All experiments are performed with 10-fold cross-validation.

Table 21. Decision tree from discretized credit approval dataset

Number of leaves	18		
Size of the tree	25		
Accuracy	87.2464%		
Confusion matrix		Predicted +	Predicted -
	Actual +	255	52
	Actual -	36	347

We'll try to remove attributes from the original dataset in the following order; A4, and A5 based on the found functional dependencies, {A4} -> {A5}, {A5} -> {A4}. Table 22 shows the property of the decision tree generated from the credit approval dataset whose attribute A4 is omitted.

Table 22. Decision tree from credit approval dataset minus A4 attribute

Number of leaves	17		
Size of the tree	24		
Accuracy	87.2464%		
		Predicted	Predicted

Confusion matrix		+	-
	Actual +	255	52
	Actual -	36	347

Table 23 shows the property of the decision tree generated from the credit approval dataset whose attribute A5 is omitted.

Table 23. Decision tree from credit approval dataset minus A5 attribute

Number of leaves	18		
Size of the tree	25		
Accuracy	87.2464%		
Confusion matrix		Predicted +	Predicted -
	Actual +	255	52
	Actual -	36	347

Next, we'll try to remove attributes from the original dataset in the following order; A10 and A11 based on the found functional dependency, {A11} -> {A10}. Table 24 shows the property of the decision tree generated from the credit approval dataset whose attribute A10 is omitted.

Table 24. Decision tree from credit approval dataset minus A10 attribute

Number of leaves	19		
Size of the tree	26		
Accuracy	86.9565%		
Confusion matrix		Predicted +	Predicted -
	Actual +	256	51
	Actual -	39	344

Table 25 shows the property of the decision tree generated from the credit approval dataset whose attribute A11 is omitted.

Table 25. Decision tree from credit approval dataset minus A11 attribute

Number of leaves	18		
Size of the tree	25		
Accuracy	87.2464%		
Confusion matrix		Predicted +	Predicted -
	Actual +	255	52
	Actual -	36	347

If we compare table 21 ~ table 25, removing the redundant attribute A4 generates some good results as shown in table 22. But, because the functional dependencies are relatively simple, we do not get much improvement in the decision trees.

All in all, we can conclude that checking functional dependency to remove redundant conditional attributes is statistically valid for the real-world datasets, and we may get a better decision tree when the found functional dependencies contain some variety of values of many-to-one relationships and the values of Pearson chi-square is relatively large as summarized in table 26.

Table 26. Pearson chi-square of attributes in the datasets of the experiment

Attribute	education, education_num in adult dataset	pdays-poutcome, previous in bank dataset	A4, A5 in credit approval dataset	A10, A11 in credit approval dataset
Pearson Chi-square	293052	45211	1380	690

5 Conclusion

When we supply training datasets for the task of data mining, independence between conditional attributes in the datasets is recommended as a preprocessing task. A well-known statistical method for the task is the chi-squared test. We can choose two categorical or nominal attributes, and we can determine their independence easily. But, if we have many attributes in the datasets, choosing some appropriate two attributes will be a challenging task. Therefore, if we have some automatic tool that can help us choose appropriate attributes to do the test, it'll be very good for us to save time in our data mining task. On the other hand, functional dependencies represent many-to-one correspondence between subsets of attributes in the relation or table, including one-to-one relationships of values that appear in a relation. So, if there is a functional dependency, it is highly probable that there is a statistical dependency between the two attributes or between the two subsets of attributes that appear between the left-hand side and the right-hand side of the functional dependency. There are several algorithms for the automated discovery of functional dependencies based on data, but there can be many fake functional dependencies because of not enough data. So, it is very natural that using the found functional dependencies with the automatic tools we want to check their statistical dependencies

of them. In this work, we have checked the statistical dependence by using three different public datasets to see their statistical significance of found functional dependencies and generated decision trees after removing dependent attributes. Experiments revealed that the found functional dependencies have statistical dependence in real-world datasets. Additionally, we could find out that decision trees generated some better results after dropping some redundant or dependent attributes, especially when found functional dependencies have a variety of values in the many-to-one relationships and the value of Pearson chi-square is relatively large for the attributes. Future works can be, for generality, the extension of FDtool that can be applied to datasets having attributes of less than 27 only.

Acknowledgement:
This work was supported by Dongseo University, "Dongseo Frontier Project" Research Fund of 2021.

References:

[1] C.J. Date. *Database Design and Relational Theory: Normal Forms and All That Jazz*, 2nd ed., Apress, 2019.

[2] N. Asghar, A. Ghenai, *Automatic Discovery of Functional Dependencies and Conditional Functional Dependencies: A Comparative Study*, University of Waterloo, April 2015.

[3] G.K. Kanji, *100 Statistical Tests*, 3rd ed., SAGE Publications Ltd, 2006.

[4] *SPSS Tutorial: Chi-square test of independence,* https://libguides.library.kent.edu/spss/chisquare, 2022.

[5] B.T. Jijo, A.M. Abdulazeez, Classification Based on Decision Tree Algorithm for Machine Learning, *Journal of Applied Science and Technology Trends*, Vol.2, No.1, 2021, pp. 20-28.

[6] M. Belkin, D. Hsu, S. Ma, S. Mandal, Reconciling modern machine-learning practice and the classical bias-variance trade-off, *PNAS*, Vol. 116, No. 32, 2019, pp. 15849-15854.

[7] P. Tare, S. Mishra, M. Lakhotia, K. Goyal, Bias Variance Trade-off in Classification Algorithms on the Census Income Dataset, *International Journal of Computer Techniques*, Vol. 6, Issue 3, 2019, pp. 1-5.

[8] M. Robnik-Sikonja, I. Kononenko, Attribute dependencies, Understandability and Split Selection in Tree-Based Models, *Proceedings of the Sixteenth International Conference on Machine Learning,* 1999, pp. 344-353.

[9] R. Elshwi, M.H. Al-Mallah, S. Sakr, On the Interpretability of Machine Learning-based Model for Predicting Hypertension, *BMC Medical Informatics, and Decision Making*, Vol.19, Article 146, 2019.

[10] J.R. Quinlan, *C4.5: Programs for Machine Learning*, Elsevier, 2014.

[11] C.J. Date, *An Introduction to Database Systems,* 8th ed., Pearson, 2003.

[12] L. Caruccio, S. Cirillo, V. Deufemia, and G. Polese, Incremental Discovery of Functional Dependencies with a Bit-vector Algorithm, *Proceedings of the 27th Italian Symposium on Advanced Database Systems*, 2019, pp. 146-157.

[13] J. Liu, J. Li, C. Liu, and Y. Chen, Discover dependencies from data – a review, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 24, No. 2, 2012, pp. 251-264.

[14] P. Bohannon, W. Fan, F. Geerts, X. Jia, A. Kementsietsidis, Conditional Functional Dependencies For Data Cleaning, *IEEE 23rd International Conference on Data Engineering*, 2007, DOI: 10.1109/ICDE.2007.367920

[15] R. Salem, A. Abdo, Fixing Rules for Data Cleaning Based on Conditional Functional Dependency, *Future Computing and Informatics Journal 1*, 2016, pp. 10-26.

[16] F. Azzalini, C. Criscuolo, L. Tanca, FAIR-DB: Functional Dependencies to Discover Data Bias, *Proceedings of the EBDT/ICDT 2021 Joint Conference*, 2021.

[17] D. Nguyen, L.T.T. Nguyen, B. Vo, W. Pedrycz, Efficient Mining of Class Association Rules with the Itemset Constraint, *Knowledge-Based Systems*, Vol.103, 2016, pp. 73-88.

[18] M. Nasr, M. Hamdy, D. Hegazy, K. Bahnasy, An Efficient Algorithm for Unique Class Association Rule Mining, *Expert Systems with Applications*, Vol. 164, 113978, 2021, https://doi.org/10.1016/j.eswa.2020.113978

[19] S. García, J. Luengo, J. Sáez, V. López, F. Herrera, A Survey of Discretization Techniques: Taxonomy and Empirical Analysis in Supervised Learning, *IEEE Transactions on Knowledge and Data*

Engineering, 2013, DOI:10.1109/TKDE.2012.35

[20] U.M. Fayyad, K.B. Irani, Multi-interval discretization of continuous-valued attributes for classification learning, *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence*, 1993, pp.1022-1027.

[21] X. Wu, V. Kumar, J.R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G.J. McLachlan, A. Ng, B. Liu, P.S. Yu, Z. Zhou, M. Steinbach, D.J. Hand, D. Steinberg, Top 10 algorithms in data mining, *Knowledge and Information System*, Vol. 14, 2008, pp. 1-37.

[22] J.N.K Rao, A.J. Scott, The Analysis of Categorical Data from Complex Sample Surveys: Chi-Squared Tests for Goodness of Fit and Independence in Two-Way Tables, *Journal of the American Statistical Association*, Vol. 76, No. 374, 1981, pp. 221-230.

[23] M. Buranosky, E. Stellnberger, E. Pfaff, D. Diaz-Sanchez, C. Ward-Caviness, *FDTool: a Python application to mine for functional dependencies and candidate keys in tabular form* [version 2, peer review: 2 approved], F1000Research 2019, 7:1667, https://doi.org/10.12688/f1000research.16483.2

[24] A. Field, *Discovering Statistics Using IBM SPSS Statistics: North American Edition*, 5th ed., SAGE Publications Ltd., 2017.

[25] E. Frank, M.A. Hall, I.H. Witten, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, Fourth Edition, 2016.

[26] Dua and C. Graff, *UCI Machine Learning Repository* [http://archive.ics.uci.edu/ml] Irvine, CA, University of California, School of Information and Computer Science, 2019.