

Weng Xu - 2203589 - Data Mining & Visualisation

by Xu Weng

Submission date: 10-Mar-2023 01:08PM (UTC+0000)

Submission ID: 199004075

File name: 40709_Xu_Weng_Weng_Xu_-_2203589_-_Data_Mining_Visualisation_482892_990029332.docx
(419.92K)

Word count: 2305

Character count: 12497

Report of Data Mining and Visualisation

Weng Xu 2203589

University of Wales Trinity Saint David,

Assignment of Data Mining & Visualisation

Contents

1. Introduction.....	1
2. Description.....	2
2.1. Datafolder.....	2
2.2. Overview of the dataset.....	3
2.3. Attribute information.....	3
2.4. Data type.....	4
2.5. Reason.....	4
2.6. Requirement on Data Preprocessing.....	5
3. Data Analysis.....	5
3.1. Which attributes most lead to a good labor relationship (good event)?.....	5
3.2. Which attributes are most likely to lead to poor labor relations (bad event)?.....	6
3.3. What is the distribution of wage increases over the three-year period?.....	6
3.4. Is there a correlation between wage increases and the cost of living allowance?.....	6
4. Summary of Suitability.....	7
4.1. Relevance to the research question.....	7
4.2. Completeness and accuracy of data.....	7
4.3. Representativeness of the data.....	7
4.4. Adequacy of the data for the analysis.....	7
4.5. Ease of data preprocessing.....	7
5. Conclusion.....	8
Video Link.....	8
References.....	8

1. Introduction

From the UCI Machine Learning Repository Center, we have the Labor Relations Data Set. Several organizations contributed to its creation, including Collective Bargaining Review, a monthly publication, and Labour Canada's Industrial Relations

Information Service[1].

Among the data are all collective agreements signed in the business and personal services sectors in Canada between 87 and 88 by locals with at least 500 workers (teachers, nurses, university staff, police, etc.). Each record in the dataset contains 16 attributes, and records with missing values have been deleted. Using the dataset, you can test a two-tier approach with positive and negative examples for learning. As a result of this report, we describe the dataset, how complete it is, and whether there are any relevant analytical issues.

2. Description

2.1. Datafolder

The datafolder contains the following files: parent directory, C4.5/, Index, labor-negotiations.data, labor-negotiations.names and labor-negotiations.test. In the dataset, each of them will have a different role to play (Figure 1).

Index of /ml/machine-learning-databases/labor-negotiations

- [Parent Directory](#)
- [C4.5/](#)
- [Index](#)
- [labor-negotiations.data](#)
- [labor-negotiations.names](#)
- [labor-negotiations.test](#)

Apache/2.4.6 (CentOS) OpenSSL/1.0.2k-fips SVN/1.7.14 Phusion_Passenger/4.0.53 mod_perl/2.0.11 Perl/v5.16.3 Server at archive.ics.uci.edu Port 443

Fig. 1. Labor Relations Data Set

Parent Directory: In this folder, all dataset-related files are stored, and it serves as the main folder for all dataset-related files within the dataset[2].

C4.5/: The C4.5 algorithm is used to train decision tree models, and the data from these models can be found in the folder on this page[3].

Index: The file contains a list of the corresponding instance numbers and datasets for all of the datasets that are part of this project.

Labor-negotiations.data: Contains a set of datasets containing all of the instances in the dataset, along with the values for each attribute found in each instance.

Labor-negotiations.names: There are a number of attributes contained in the file, including the name, type, and value of the attributes of the dataset.

Labor-negotiations.test: There is a file in this package that contains a portion of the test dataset used to assess the model's accuracy and generalizability.

2.2. Overview of the dataset

An example of the dataset is 57 instances of real data that was donated to the community in 1988. There are 16 attributes in the dataset, which include categorical, integer, and real data. In addition, there are no missing values in the dataset, nor is any task associated with it specified. During the past year, the dataset has received a total of 100,859 web hits (Table 1).

Table.1. Data Abstract

Abstract: From Collective Bargaining Review

Data Set Characteristics:	Multivariate	Number of Instances:	57	Area:	Social
Attribute Characteristics:	Categorical, Integer, Real	Number of Attributes:	16	Date Donated	1988-11-01
Associated Tasks:	N/A	Missing Values?	No	Number of Web Hits:	100889

2.3. Attribute information

For each of the 16 attributes in the dataset, there are four categories to be considered:

1. It is worth noting that there are several categories of salary and allowances available to employees, including the yearly salary increase, the living expenses allowance, the standby allowance, shift allowance, and educational allowances.

2. In order to qualify for the working hours and leave category, you must be in compliance with all of the following: Duration of employment agreement, working hours, statutory holidays, and vacation days.

3. A pension plan, a health insurance plan, and plans that cover employer contributions to health care benefits.

4. Among the others, there are employers who provide assistance to long-term disabled employees, employers who contribute financially for aftercare services, and many others.

Therefore, users will be able to understand various attributes and characteristics contained in this dataset better and will therefore be able to conduct further analyses and research with more precision[4] (Table 2).

Table.2. ARFF Header

No.	Attribute	Description	Value Range
1	duration of agreement	Duration of the labor agreement	
2	wage1.wage	Wage increase in first year of contract	[2.0 .. 7.0]
3	wage2.wage	Wage increase in second year of contract	[2.0 .. 7.0]
4	wage3.wage	Wage increase in third year of contract	[2.0 .. 7.0]
5	cola	Cost of living allowance	[none, tcf, tc]
6	hours.hrs	Number of working hours during the week	[35 .. 40]
7	pension	Employer contributions to pension plan	[none, ret_allw, empl_contr]
8	stby_pay	Standby pay	[2 .. 25]
9	shift_diff	Shift differential - supplement for work on II and III shift	[1 .. 25]
10	educ_allw.boolean	Education allowance	[true false]
11	holidays	Number of statutory holidays	[9 .. 15]
12	vacation	Number of paid vacation days	[ba, avg, gnr]
13	lngtrm_disabil.boolean	Employer's help during employee long-term disability	[true , false]
14	dntl_ins	Employers contribution towards the dental plan	[none, half, full]
15	bereavement.boolean	Employer's financial contribution towards the covering the costs of bereavement	[true , false]
16	empl_hplan	Employer's contribution towards the health plan	[none, half, full]

2.4. Data type

A variety of attributes are offered in this package, including integers, real numbers, Booleans, and various value sets, depending on the attribute type. The values of some attributes can be discrete, while other values can be continuous. In addition to helping you make more informed decisions regarding labor negotiation factors, such as salary, allowances, and benefits, they are also useful when analyzing various labor negotiation factors[5].

2.5. Reason

Following consideration of its social relevance, multivariate nature, and labeled data availability, we chose this dataset.

Social relevance: There has been a lot of interest in labor negotiations, and I believe that this dataset could be useful for employment lawyers and labor relations professionals in the future[6].

Multivariate nature: This dataset has 16 attributes, including categorical and numeric variables, which means it can be used to test and evaluate a variety of statistical models and algorithms for machine learning.

Availability of labelled data: The dataset included labeled examples of both "good" events related to labor negotiations and "bad" events related to labor negotiations which would be useful for the development and testing of predictive models or decision support systems.

2.6. Requirement on Data Preprocessing

In all cases, the information provided does not include missing values, or attributes that have been converted into different types of data unless otherwise stated. Information provided is limited only to domain types and possible attributes, as well as examples of good and bad events, based on the provided information (Table 2).

3. Data Analysis

This data set can be used to answer the following core question: Which attributes are most likely to contribute to good labor relations (good events) or which attributes are most likely to contribute to bad labor relations (bad events); also, a number of more diffuse issues may be taken into consideration, including "How does the cost of living allowance correlate with wage increases over the period of three years?" or "What is the distribution of wage increases over a period of three years?". In some situations, you may find that choosing questions from a broader range of subjects allows you to find unexpected insights[7].

3.1. Which attributes most lead to a good labor relationship (good event)?

Association rule mining can help identify a common set of itemsets and association rules found in the good events data. The attributes are ranked based on their support and confidence values. The attributes that are most likely to lead to good labor relations are those that have the highest support and confidence values[8].

There would be a set of frequently occurring itemsets and associations rules that would emerge from the association rule mining. These rules could provide insight into which attributes are most strongly associated with good labor relations and in turn provide information about how to improve them.

3.2. Which attributes are most likely to lead to poor labor relations (bad event)?

The missing values should be removed from the categorical variables, and all binary variables should be transformed into dummy variables[9].

Through association rule mining, we can identify the frequent items and association rules present in the bad events data that occur on a regular basis. Based on the strength of the associations, we can then rank those attributes based on their support and confidence values. It is evident from the highest support and confidence values that the attributes are more likely to result in a poor labor relations situation.

Among the outcomes of the association rule mining would be a set of frequent itemsets, as well as associations rules, which would be a source of insights as to which of the attributes is most strongly associated with bad labor relations.

3.3. What is the distribution of wage increases over the three-year period?

In the data mining process, descriptive statistics are used[10].

In order to analyze the wage increases for the three years in the medial direction, we can calculate the mean, median, mode, and standard deviation of the wage increases for each year as well as produce a histogram or boxplot that illustrates the distribution of wage increases over the three years in the medial direction.

The model would not be generated because this is a descriptive analysis rather than a forecast question.

3.4. Is there a correlation between wage increases and the cost of living allowance?

Using correlation analysis to mine data[11]. The missing values should be removed from the categorical variables, and all binary variables should be transformed into dummy variables.

This process of analysis can be accomplished using Pearson's correlation to calculate the correlation coefficient between wage increases and cost of living allowances. We can also visualize this relationship with the creation of a scatter plot to help you understand the relationship.

In order to model the correlation between two variables from the data that is received from the correlation analysis, one would have to find a correlation coefficient value that indicates the strength and direction of the correlation between the two

variables.

4. Summary of Suitability

It is possible to use several criteria to evaluate data within the "work relation" dataset in different ways.

4.1. Relevance to the research question

Since this study is relevant since the dataset contains information about work-related attributes such as wages, hours, pensions, standby pay, shift differentials, shift differentials between shifts, holidays, vacations, and insurance policies, it is relevant to the study[12].

4.2. Completeness and accuracy of data

Generally speaking, the data sets seem to be complete, as there seem to be no missing values in any of the variables. However, it may not be possible from what is provided to determine the accuracy of the data.

4.3. Representativeness of the data

Further information about the sampling method and the source of the data would be helpful to determine whether the dataset represents the overall population or a specific industry. Further information is needed about the sampling method and the source of the data.

4.4. Adequacy of the data for the analysis

The dataset contains all the variables necessary to answer the research question, such as wages, hours, and pensions. However, in order to analyze the research question in a more detailed manner, further information about the variables is needed.

4.5. Ease of data preprocessing

The dataset seems to be straightforward and easy to preprocess since there is no matter of missing values or outliers in the dataset.

According to the provided information, it appears that the "work relation" dataset may be capable of being analyzed, however, more information is needed about the

representativeness of the data and the definitions of some of the variables.

5. Conclusions

For the analysis of the "Work Relation" dataset, a variety of techniques were used, including the cleaning and preprocessing of the data, and the application of data mining tools. As a result of the dataset, many different attributes were included, including duration, salary, cost of living allowance, pensions, standby pay, differentials between shifts, holidays, vacations, dental insurance, and employee health plans.

It was the most challenging part of the analysis for each of the research questions to select an appropriate data mining technique[9]. In order to select the most effective and accurate data mining and statistical analysis method, it was essential that one had a thorough understanding of the dataset and data mining techniques.

In general, the study had a number of limitations including the small sample size and the fact that only a three-year period of time was covered in the dataset. It is possible, then, that the dataset could not provide a fuller understanding of long-term labor relations if it covered a longer period of time.

A thorough understanding of data mining techniques and statistical analysis was required in order to analyze the "Work Relation" dataset. It is unfortunate that the sample size was so small and the study was conducted within a limited period of time. The results from the analysis, however, can have a significant impact on labor relations in the future and may serve as a guide for future studies.

Video Link

https://youtu.be/_L8oPm4J8e8

References

- [1] UCI Machine Learning Repository: Labor Relations Data Set. (n.d).
<https://archive.ics.uci.edu/ml/datasets/Labor+Relations>
- [2] Molnar, C. (2020). Interpretable Machine Learning. Lulu.com.
- [3] Mantas, C. J., & Abellán, J. (2014). Credal-C4.5: Decision tree based on imprecise probabilities to classify noisy data. Expert Systems With Applications, 41(10), 4625–4637. <https://doi.org/10.1016/j.eswa.2014.01.017>

- [4] Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A. K., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Stei-ner, B., Tucker, P. G., Vasudevan, V. K., Warden, P. Zheng, X. (2016). TensorFlow: a system for large-scale machine learning. *Operating Systems Design and Implementation*, 265–283. <https://doi.org/10.5555/3026877.3026899>
- [5] Mueller, J. P., & Massaron, L. (2021). *Machine Learning For Dummies*. John Wiley & Sons.
- [6] Deisenroth, M. P., Faisal, A. A., & Ong, C. S. (2020). *Mathematics for Machine Learning*. Cambridge University Press.
- [7] Ghayoumi, M. (2021). *Deep Learning in Practice*. CRC Press.
- [8] Han, J., Pei, J., & Kamber, M. (2011). *Data Mining: Concepts and Techniques*. Elsevier.
- [9] Linoff, G. S., & Berry, M. J. A. (2011). *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management*. John Wiley & Sons.
- [10] Olson, D. L., & Delen, D. (2008). *Advanced Data Mining Techniques*. Springer Science & Business Media.
- [11] Otero, P., Scott, P., & Martin, S. (2022). *MEDINFO 2021: One World, One Health — Global Partnership for Digital Innovation: Proceedings of the 18th World Congress on Medical and Health Informatics*. IOS Press.
- [12] Zaki, M. J., & Meira, W., Jr. (2019a). *Data Mining and Machine Learning: Fundamental Concepts and Algorithms*. Cambridge University Press.

FINAL GRADE

GENERAL COMMENTS

Instructor

78/100

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6

PAGE 7

PAGE 8

PAGE 9

GRADING FORM: DATA MINING/ VISUALISATION

XU WENG

78

KNOWLEDGE AND UNDERSTANDING

Dataset Suitability (What features of the dataset lend themselves to data mining? Has it been mined previously? Why does the student feel that they have the contextual background to work with this data?)



Fair knowledge of the Labour dataset and its inherent features are explained in great detail.

27

COGNITIVE AND INTELLECTUAL

Objectives (What does the student plan to investigate/expect to find? Are these SMART?)



Each proposed question for future analysis can be considered specific, measurable, achievable, relevant and time-bound.

8

PRACTICAL APPLICATION SKILLS

Dataset Description (Where did the data come from? What does the data describe? Is the sample large enough? What data formats and types are involved?)



Dataset description, location, types and sample validity are all addressed sufficiently well in submission.

37

TRANSFERABLE SKILLS FOR LIFE

Presentation/Slides (Clarity, readability, flow, accuracy, completeness, performance)



Fairly well organised document that has incorporated all assessment elements to a great

6

extent. Word count is however on the high side.