International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

Innovative Data Mining Techniques for Healthcare and Social Sciences

Ankita Moreshwar Itankar[1], Vijaya Kamble[2]

[2]Assistant Professor, Department of Computer Science and Engineering, GuruNanak Institution & Technology, Kalmeshwar Road, Nagpur, India

[3]Assistant Professor, Department of Computer Science and Engineering, GuruNanak Institute of Engineering & Technology, Kalmeshwar Road, Nagpur, India

Abstract: ***Data mining is an analytical technique used to discover relationships between variables and find patterns in data. Using these findings, data mining can create predictive models (e. g. target variable forecasting, label classification) or identify different groups in data (e. g., clustering). Although databases are mature and widely used in many fields, including computer vision, natural language processing, and bioinformatics, databases have only recently been widely used in the social and medical sciences. In fact, there is an interest in developing data mining techniques suited to specific exploratory problems that arise in many fields, including the social sciences (Attewell et al., 2015). An important problem in the social sciences is to identify the factors that encourage or inhibit population growth; knowledge is the single best tool for this problem. Identifying these factors is important for planning good public policy and allocating housing resources based on future population growth. To understand and explain population growth in the context of its fundamental principles (for example, economic, social, architectural, or material impact), researchers use statistical methods such as cross- sectional analysis (Carlino and Mills 1987; Clark and Murphy 1996; Beeson et al.2001; Chi and Voss 2010; Chi and Marcouiller 2011; Iceland et al.2013). However, these studies sometimes show conflicting results due to multiple variables (a close linear relationship between two or more variables). More specifically, these previous studies included strategies that did not consider the success of input devices.***

1. Introduction

In the social sciences, a very important problem is that of identifying the factors that promote or hinder population growth; data mining tools are ideal for addressing this problem. Identification of such factors is important for the effective public policy development plan and the allocation of infrastructure investments that align with the future population growth.

In the healthcare sciences, a very important problem is that of determining the acceptance/ rejection of cancer treatment plans; data mining tools are ideal for addressing this problem. For example, proposed radiation therapy (RT) plans need to be reviewed by RT experts to determine whether these RT plans are acceptable. This review process involves a laborious manual evaluation and a large amount of human resources. Thus, an automated system to classify the proposed RT plans as acceptable or erroneous can be useful in reducing the overload of RT experts and eliminating human errors.

2. Objectives

1) The principal objective of this dissertation was to develop data mining algorithms that outperform conventional data mining techniques on social and healthcare sciences.
2) Toward this objective, this dissertation developed two data mining techniques, each of which addressed the limitations of a conventional data mining technique when applied in these contexts.
3) To propose a novel data mining methodology that can identify significant input factors affecting a given target variable, even in the presence of multicollinearity.
4) To propose method can rank these input factors according to their influence on the target variable.
5) To apply our proposed method to a real dataset in demographic research identification of significant factors promoting or hindering population growth.

3. Proposed Plan Work

The following steps describe how the planning process combines CART and Cohen's d to identify trade - offs/related effects for different purposes.

1) Divide cities into different groups using the CART algorithm.
2) Put the cities in the two groups with the highest goal difference into a group. Therefore, these groups will have cities with high target values and variable values. Again, form a group by taking cities with the lowest values from the two groups. These groups are called the super group and subgroup, respectively.
3) For each variable, calculate the Cohn d - index of the upper and lower groups.
4) Rank the difference by Cohen's d index; those with the highest (as the lowest) index are the variables/features that have the highest (as the lowest) effect on the target variable.

Figure 1 shows the planning process, the decision tree with Cohen's d - index. Note that Cohen's d for each variable is an indicator of the effect level of the difference between the groups measuring it. Therefore, when Cohen's d measures the effect of each variable on population growth, the relationship between different strategies does not affect the calculation.

The general idea behind the above process is as follows. The upper and lower groups have cities with different input

values this is characteristic of the group obtained using the CART algorithm. In addition, in the upper and lower groups, there are cities with high and low values of the target variable, respectively. Since the proposed method uses Cohen's d to find input variables for which these two groups differ significantly, it is necessary to subtract the input variable with the highest (lowest) Cohen's d prime/factors, regardless of the relationship between the input variables. Has the highest (lowest) effect on the target variable. Note that another way to find the CART group of cities in the upper (lower) group is to include all the cities with the highest (lowest) value of the target variable. However, it is better to use CART groups to find upper and lower groups because groups of CART groups are homogeneous across different value graphs, but also across different devices.

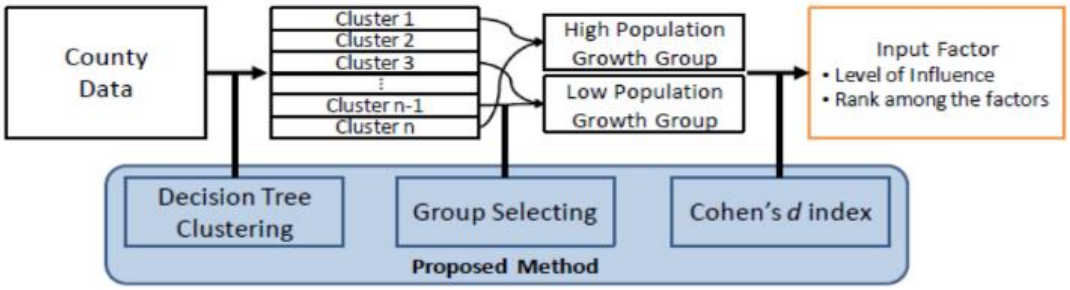

Figure 1: Process of the proposed method, decision tree combined with Cohen's d

4. Conclusions

This section presents the results obtained by applying the proposed method to actual city - level US data described in Section 2.3. Using the CART algorithm and population growth as the target variable, we divided 3, 108 cities into 10 groups as shown in Figure 2. As mentioned earlier, the decision tree structure creates a simply defined group. In particular, given a city, it is simple to determine the category it belongs to: starting with and following the basis of the given city - provided sorting rules up to the page/group. For example, Figure 2 shows that group 10 with the largest average population in cities (21: 57%) has cities where the average income is greater than or equal to about $42; 447: 5 and the average temperature in January is 31: 250F. On the other hand, there are Group 2, the group with the smallest population in the middle (6: 25%), provinces with an average income of less than $42; 447: 5, population less than 8: 25 people/square mile, January to July temperature greater than or equal to 0: 130F. The number above each node denotes the number of counties inside the node; the label on an interior node is the variable chosen to partition the data records in this node into its two child nodes; the splitting rule and splitting point are given on the edge from the parent node to each child node; the number inside each lea f node denotes the target variable's average value of the data records in the cluster.

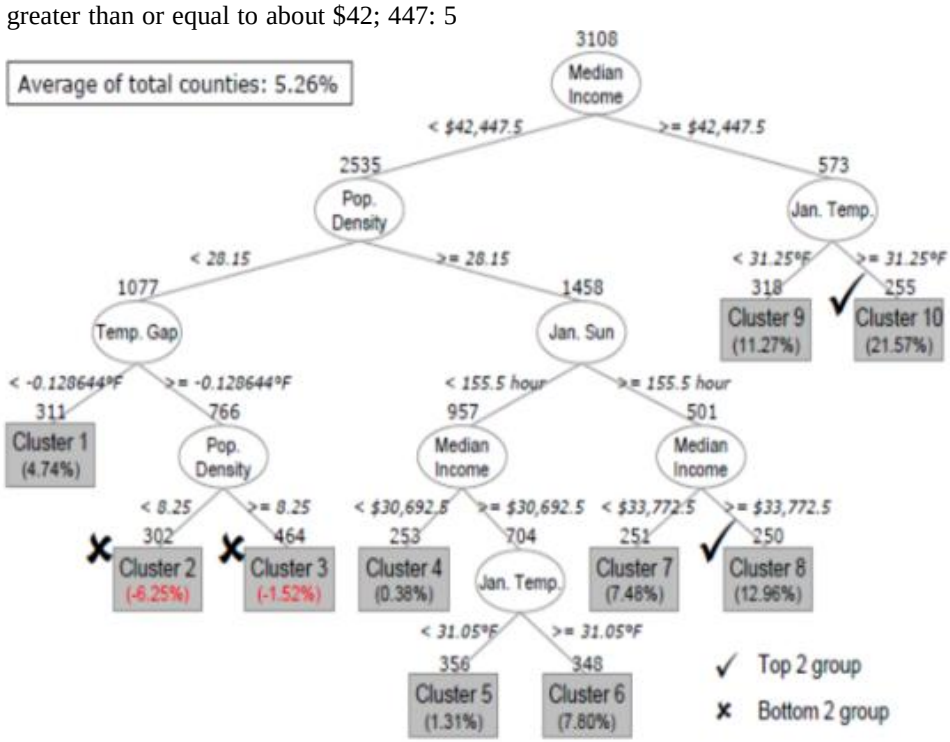

Figure 2: Decision tree obtained by applying CART to the county-level dataset.

References

[1] Paul Attewell, David B. Monaghan, and Darren Kwong. Data Mining for the Social Sciences: An Introduction. University of California Press, 2015.
[2] Francis R. Bach, David Heckerman, and Eric Horvitz. Considering cost asymmetry in learning classifiers. The Journal of Machine Learning Research, 7: 1713–1741, 2006.
[3] Patricia E. Beeson, David N. DeJong, and Werner Troesken. Population growth in US counties, 1840–1990. Regional Science and Urban Economics, 31 (6): 669–699, 2001.
[4] Andrew P. Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition, 30 (7): 1145–1159, 1997.
[5] Leo Breiman, Jerome Friedman, Charles J. Stone, and Richard A. Olshen. Classification and regression trees. Chapman & Hall/CRC, 1984.
[6] Carla E. Brodley and Mark A. Friedl. Identifying mislabeled training data. Journal of Artificial Intelligence Research, 11: 131–167, 1999.
[7] David L. Brown. Migration and community: Social networks in a multilevel world. Rural Sociology, 67 (1): 1–23, 2002.
[8] David L. Brown, Glenn V. Fuguitt, Tun B. Heaton, and SabaWaseem. Continuities in size of place preferences in the united states, 1972–1992. Rural Sociology, 62 (4): 408–428, 1997.

Author Profile

Ankita Moreshwar Itankar, M- Tech final year in Computer Science & Engineering from GuruNanak, Institute of Engineering & Technology, Nagpur

Prof Vijaya Kamble is Assistant Professor in Guru Nanak Institute of Engineering & Technology