IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 2, June 2023

Advanced Data Mining Techniques for Medicinal and Societal Sciences

Ms. Swati G. Bansod and Mr. Gurudev B Sawarkar
Department of Computer Science & Engineering
Wainganga College of Engineering and Management, Dongargaon, Nagpur, Maharashtra, India

Abstract: *This paper centers around advancement of information mining calculations that outflank traditional information mining methods on friendly and medical care sciences. Toward this goal, this exposition creates two information mining methods, every one of which tends to the impediments of a traditional information mining strategy when applied in these specific circumstances. To start with, we propose an original information mining system that can recognize critical information factors influencing a given objective variable, even within the sight of multicollinearity. Additionally, the proposed technique can rank these information factors as per their impact on the objective variable. Then, we apply our proposed technique to a genuine dataset in segment research ID of huge variables advancing or upsetting populace development (Part I). Secondlly, we foster a characterization technique for imbalanced information where the greater part class has essentially a bigger number of occasions than the minority class. Then, at that point, we apply our proposed imbalanced-information arrangement technique to eleven open datasets, the vast majority of them connected with medical services sciences (Part II).*

Keywords: *Data Mining, algorithm, social sciences, healthcare sciences*

I. INTRODUCTION

Data mining is an analytic process for discovering systematic relationships between variables and for finding patterns in data. Using those findings, data mining can create predictive models (e.g., target variable forecasting, label classification) or identify different groups within data (e.g., clustering). Although data mining is already well-established and widely used in many fields including computer vision, natural language processing, and bioinformatics, data mining techiniques were not as widely used in the social and healthcare sciences until recently. Indeed, there is a growing interest to develop data mining techniques specifically tailored for the unique discovery problems arising in many fields such as the social sciences (Attewell et al., 2015). 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. To understand and explain population growth in terms of its underlying factors (i.e., economic, social, infrastructural, or amenity factors), population researchers have used statistical models such as linear regression analyses (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 showed inconsistent results between one another due to the presence of multicollinearity — a near-linear relationship between two or more input factors. Specifically, these previous studies included input factors without considering the statistical dependence among the included input factors.

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. However, an RT-plan classification system developed using conventional classification methods would have poor erroneous-case detection performance. This is because (1) among the RT plans, erroneous cases are very rare

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 2, June 2023

and (2) conventional classification methods are designed to minimize the number of misclassified cases over the training data, and thus they would tend to predict the vast majority (if not all) of the test set cases as acceptable cases.

This paper focuses on development of data mining algorithms that outperform conventional data mining techniques on social and healthcare sciences. Toward this objective, this dissertation develops two data mining techniques, each of which addresses the limitations of a conventional data mining technique when applied in these contexts. First, we propose a novel data mining methodology that can identify significant input factors affecting a given target variable, even in the presence of multicollinearity. Moreover, the proposed method can rank these input factors according to their influence on the target variable. Then, we apply our proposed method to a real dataset in demographic research identification of significant factors promoting or hindering population growth (Part I). Second, we develop a classification method for imbalanced data-data where the majority class has significantly more instances than the minority class. Then, we apply our proposed imbalanced-data classification method to eleven open datasets, most of them related to healthcare sciences (Part II).

II. REVIEW OF LITERATURE

Community researchers using secondary data draw usually from two approaches to explain community growth. The first approach, characteristic of very early studies of community growth, focused on understanding community growth from an economic or a demographic perspective independently. Using this approach, academics with particular training were studying community growth based mainly on their area of expertise. Typically, economists were measuring economic growth through economic data while demographers and sociologists were examining community growth as measured by demographic data (see, for example, Pearl and Reed 1920; Pritchett 1891).

The second approach is more comprehensive and uses different types of information (e.g., demographic, economic, environmental, and policy variables) to explain community growth. As research advanced, researchers examining economic growth realized that demographic factors (e.g., population density, percentage of minorities present, educational attainment of the population), environmental factors (e.g., climate, topography, natural amenities), and policy factors (e.g., taxes, subsidies, regulations) needed to be included as input factors in their models in addition to economic factors (for example, Carlino and Mills 1987; Clark and Murphy 1996; Quigley 1998; Deller et al. 2001). Similarly, studies examining population growth also noted the importance of combining different types of explanatory factors such as economic factors (e.g., income, labor mobility), and cultural and environmental factors (e.g., personal preferences on community and residential characteristics) as predictors of population growth besides demographic factors (for example, Leslie and Richardson 1961; Sjaastad 1962; Golant 1971; Zelinsky 1971; Speare 1974; Fuguitt and Zuiches 1975; Greenwood 1975; Carlino and Mills 1987; Clark and Murphy 1996; Brown et al. 1997; McGranahan 1999; Deller et al. 2001; Beeson et al. 2001; Rupasingha and Goetz 2004; Brown 2002).

Recently, some studies have focused on improving community research models by overcoming the issue of multicollinearity. The issue of multicollinearity arises when there is a near-linear relationship among two or more input variables, and this multicollinear- ity leads to inaccurate estimates or low statistical significance values. The favored two stage least squares lagged adjusted regressions of the 1990s were very vulnerable to multicollinearity. Previous studies that used regression analyses selected several input variables in the same type of category (i.e., high school degree ratio and college degree ratio in the education category) without considering the statistical dependence from other variables and thus, are most likely exposed to the risk of multicollinearity. When multicollinearity is predominant, (1) small changes in the data produce wide swings in the parameter estimates; (2) coefficients may have very high standard errors and low significance levels even though they are jointly significant and the R2 for the regression is quite high; and, (3) coefficients may have the wrong sign or implausible magnitude in a regression analysis (Greene, 2012). Therefore, to overcome this multicollinearity issue, some researchers used only a subset of the input variables for calculating the level of significance (see, for example, Chi and Voss 2010; Iceland et al. 2013). Alternatively, Chi and Marcouiller (2011) and Deller et al. (2001) overcame this problem by merging the input variables into several category variables using Principal Factor Analysis (PFA) and Principal Component Analysis (PCA) respectively.

Table 1 shows the list of significant factors for population growth determined by previous regression-based studies. One observation, as mentioned above, is that previous studies did not provide the level of influence of each input variable on

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 2, June 2023

population growth. Another important observation of this table is that the results of previous population growth studies are not consistent with each other studies.

Table 1: List of significant factors for population growth

Variable	Carlino et.al.	Clark et.al.	Beeson et.al.	McGranahan et.al.	Chi. et.al.	Significance Ratio (%)
Median Income	@	@		#	@	100
College Ratio		@		@		67
Temperature Gap		@		#		100
Poverty Ratio		#		@		0
Asian Ratio			#			100
Water Area Ratio			@	@		100
Highway	@	#			#	33
Black Ratio	#	@			@	67
Population Density				@	@	100
January Sun		@		@		100
Local Net	#	#				0
Employment Rate	@	@				100
Hispanic Ratio						100

Note:

@-denotes the variables which were determined as significant for population growth by the corresponding regression analysis.

# denotes the variables which were determined as non-significant for population growth by the corresponding regression analysis.

Unmarked cells indicate that the corresponding study did not include the corresponding variable

III. METHODOLOGY

Communities often face significant economic and social challenges that must be understood and overcome to ensure a stable and sustainable setting for their inhabitants and the physical environment where they reside. As communities constantly change, understanding the factors that promote such change and the consequences of such change is critical. For instance, in the case of communities with an initially low population density experiencing boomtown scenarios, examples of such factors and consequences would be physical infrastructure failing to meet the expansion demand, public policy inhibiting/limiting growth, poor social integration, and involvement in community affairs (Graber, 1974; Gilmore, 1976; Hunter and Smith, 2002; Smith et al., 2001). Without an appropriate understanding of the causes of community change, the resulting local experiences can be detrimental to the local living conditions; that, on occasion, can lead to the collapse of the community.

In order to develop integrated models capable of relating economic, policy, and geographic factors together to identify factors predicting population growth, previous studies have typically used statistical regression analyses such as ordinary least squares models or two–stage least squares lagged adjustment models (see, for example, Carlino and Mills 1987; Clark and Murphy 1996; Beeson et al. 2001; Chi and Voss 2010; Chi and Marcouiller 2011; Iceland et al. 2013). While highly important, these methodologies contain certain weaknesses. First, these statistical approaches do not determine the level of influence (importance) that each input factor has on population growth. In other words, these studies focused on identifying which input factors are better at predicting population growth, but did not rank the input factors according to their level of influence on population growth. This is because a low p-value (e.g., < 0:05) indicates that we can reject the null hypothesis (i.e., the coefficient of the corresponding input factor is equal to zero) but does not indicate the level of influence of the factor on population growth. Second, multi collinearity, which refers to a linear relationship between two or more input factors, may impact the usefulness of regression analysis (Greene, 2012; Chatterjee and Hadi, 2006; Montgomery et al., 2012). Since most previous studies selected input variables without considering their statistical dependence from each other (except the studies which introduced statistical techniques to avoid multicollinearity —

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 2, June 2023

see, for example, Deller et al. 2001; Chi and Voss 2010; Chi and Marcouiller 2011; Iceland et al. 2013), most previous studies exhibit multicollinearity between input variables and thus, this multicollinearity impacts the consistency of the results obtained using regression analysis.

To overcome the issues explained above, we develop a comprehensive data mining analysis of population growth. In this study, the proposed method employs population growth as our target variable. First, the proposed method uses decision tree clustering to group communities into several clusters so that each cluster has similar values in the target variable (i.e., population growth) and also has similar values in each input factor. This clustering allows us to find the clusters with the highest and lowest population growth and ensures that the constituents within each cluster have similar characteristics. Second, Cohen's d index is used to identify the level of influence that each input factor has on population growth by measuring the level of difference of each input factor between the clusters with the highest and lowest population growth. Even in the presence of multicollinearity, the final output of the proposed model is not affected by the correlation between input factors because decision tree clustering is not affected by the correlation between input factors and because the level of influence of the input factors on the target variable is measured independently for each input factor.

Steps

The following steps describe how the proposed method combines CART and Cohen's d to determine the level of influence of each input variable/factor on the target variable.

Use the CART algorithm to cluster the counties into several clusters.

Take the counties in the two clusters with the highest average target variable value and create a group. Thus, this group will contain counties with both a high average target value and relatively homogeneous input-variable values. Similarly, take the counties in the two clusters with the lowest average target value and create a group. These groups are referred to as a top group and a bottom group respectively.

For each input variable, calculate the Cohen's d index between the top and bottom groups.

Rank the variables according to Cohen's d index; those with the highest (respectively the lowest) index are the variables/factors with the highest (respectively the lowest) influence on the target variable.

Figure 1 illustrates the process of the proposed method, decision tree combined with Cohen's d index. Note that Cohen's d, the proposed index for measuring the level of influence of the variable between the groups, is measured independently for each variable. Therefore, when Cohen's d measures the level of influence of each input variable on population growth, the correlation between input variables does not affect the calculation.

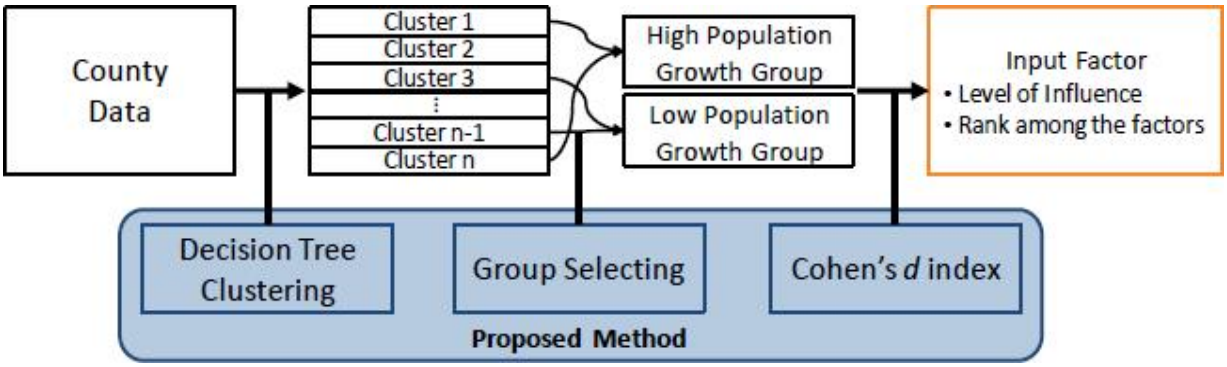

Fig. 1 Process of the proposed method, decision tree combined with Cohen's d

The general idea behind the above procedure is as follows. The top and bottom groups contain counties with relatively homogeneous input-variable values (this is a property of the clustering obtained with CART algorithm). Moreover, the top and bottom groups contain counties with high and low target variable values respectively. Since the proposed procedure uses Cohen's d to find the input variables on which these two groups differ significantly regardless of correlations between the input variables, it is reasonable to infer that the input variables/factors with the highest (lowest) Cohen's d index are those with the highest (lowest) influence on the target variable. Note that an alternative procedure to CART clustering to find the counties in the top (bottom) group is to include the individual counties with

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 2, June 2023

the highest (lowest) target variable value. However, using CART clustering to find the top and bottom groups is a better procedure because the groups clustered by CART will be homogeneous not only in the target variable values, but also in the input variable values.

IV. ADVANTAGES

The results obtained with the proposed method complement the population growth this literature in several ways.

1. Even in the presence of multicollinearity, the final output of the proposed model is not affected by the correlation between input factors because decision tree clustering is not affected by the correlation between input factors and because the level of influence of the input factors on the target variable is measured independently for each input factor.
2. The proposed method not only identifies significant factors for population growth, but also allows us to measure the level of influence that each input factor has on the target variable.

V. CONCLUSION

To improve classification performance in handling two-class imbalanced data, GU–SVM, a new imbalanced-data classification method will be proposed. The take-away message from this investigation can be summarized as follows:

- Outlier-detection and removal from both classes is crucial for handling imbalanced data.
- In fact, it makes a greater impact if one can identify and remove outliers in the minority class.
- Researchers understand the importance of selecting representative subsets of data while under sampling the majority class but how to best attain that goal is still under debate.
- The normalized-cut base approach, aiming at spreading out the majority samples evenly, provides a new angle of looking at the problem and produces competitive results.

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.
[9].Gavin Brown, Jeremy Wyatt, Rachel Harris, and Xin Yao. Diversity creation methods: A survey and categorisation. Information Fusion, 6(1):5–20, 2005.
[10]. Eunshin Byon, Abhishek K. Shrivastava, and Yu Ding. A classification procedure for highly imbalanced class sizes. IIE Transactions, 42(4):288–303, 2010.
[11]. Gerald A. Carlino and Edwin S. Mills. The determinants of county growth. Journal of Regional Science, 27(1):39–54, 1987.
[12]. Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1–27:27, 2011.
[13]. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 2, June 2023

[14]. Samprit Chatterjee and Ali S. Hadi. Regression analysis by example. Wiley- Interscience, 2006.
[15]. Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16:321–357, 2002.
[16]. Nitesh V. Chawla, Nathalie Japkowicz, and Aleksander Kotcz. Editorial: Special issue on learning from imbalanced data sets. ACM SIGKDD Explorations Newsletter, 6(1):1–6, 2004.
[17]. Guangqing Chi and David W. Marcouiller. Isolating the effect of natural amenities on population change at the local level. Regional Studies, 45(4):491–505, 2011.