ANALYSIS OF ATTRIBUTES TO IDENTIFY OPINION LEADERSHIP IN COMMUNITIES USING DECISION TREE TECHNIQUE FOR IMBALANCED DATA

Sudarat Sangkeaw[1], Piyawan Siripraseotsin[2] and Marut Buranarach[3]
[1,2]Faculty of Business Administration, Maejo University,
Sansai, Chiangmai, 50290, Thailand
[3]Language and Semantic Technology Laboratory,
National Electronics and Computer Technology Center (NECTEC), Thailand.

ABSTRACT

Opinion Leader is a key person to the Word-of-Mouth strategy in the marketing field. It represents an individual who has an ability to change the other's opinion or behavior by using an interpersonal influence. Although opinion leaders have been explored in various contexts, research on tools to identify opinion leaders for marketers are still limited. This research thus proposes to study on the attribute that can help to identify opinion leaders in community for developing a recommender system. A self-assessment questionnaire was designed and tested for validity and reliability using Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA) before used to collect data in a cyclist community. The result showed features for predicting opinion leaders are Knowledge, Extrovert and Self-confidence. Moreover, the balanced dataset with SMOTE (Synthetic Minority Over-sampling Technique) and Spread Subsampling is more effective in improving the performance of prediction results over an imbalanced dataset compared to other techniques.

KEYWORDS: Opinion Leader, Decision Tree, Imbalanced dataset, SMOTE, HDDT.

1. Introduction

It is well known that marketing strategy is usually driven by the customers' needs. Therefore, the study of consumer behavior is highly important in the marketing field. One of the theories of consumers posit is that customers can influence other customers by means of information exchange among them. These behaviors naturally occur when people are willing to help each other and share their experiences without profit. Marketers cannot control

such behavior, but recognize its huge impact. As a result, they develop a strategy called "Word-of-Mouth" which aimed at a group of people who have high reliability to be messengers for the targeted customers. Choosing right messengers are very important in influencing the other customers.

A study of opinion leaders was started from Katz & Lazarsfeld's research on the influence of mass media to the decision making of the political audiences. This led to a concept called "The Two Step Flow", which demonstrated that the message does not flow directly from press to the citizen. They discovered that it was a group of individuals who got the message and then forwarded to many people and used interpersonal influence to affect the decision making of others [1].

Many scholars studied about the behaviors and characteristics for identifying opinion leaders in a group of people. In order to identify opinion leaders in a community, there are varieties of methods, including judge's rating, voting or sociometric, self-assessment, snowballing and expert identification. Choosing an appropriate method is depending on the purposes and types of research [2]. However, these methods usually have high cost and are time-consuming.

Thus, this research proposes to find more efficient ways to identify opinion leaders by learning from the answer patterns on the questionnaire. In machine learning, the collection of data used in the learning process is very important. So we use two approaches of self-assessment combining with voting to reduce bias. The Decision Tree algorithm is used to analyze the patterns of respondents, because their results are easy to interpret and can be used for creating rules for developing an opinion leader recommender system.

The paper is organized as follows. The following sections are literature reviews related to opinion leaders and decision tree algorithm, methodology of the research, results and conclusion.

2. Literature reviews

2.1 Opinion leader concept

The concept of opinion leaders started from research of mass media communication including newspapers, magazines, radio, television, films and internet. They are mediums that can spread information from individuals to a larger community. The research of Paul

Lazarsfeld and Elihu Katz and his colleagues, during 1940-1950, studied on how the media influenced the intention or the decisions of the voters in the US presidential election. They believed that mass media is the most important factor due to the coverage of communication channels [1]. However, contrary to this hypothesis, they found that the flow of information which started from radio or press was received by a minority group of people and then forwarded to public. This group of people called “Opinion Leader” refers to the individuals in society which influenced other people in specific situations. They give comments or recommendations in an unofficial manner. This opinion leadership appears generally in society, whether at home, work, school, or wherever there is a close relationship occurred. Some opinion leaders may be influential in many topics, although they often influence on specific topics such as fashion apparel, shopping, electoral politics and sports, etc.

Further research has found that opinion leaders can influence the diffusion of innovations, according to the trend of technology adoption, which would begin with adoption by a small group of opinion leaders. The behavior of this group was important to the adoption rate of innovation in the S-shaped Curve Distribution [3]. Also studies on medical practices as opinion leader for doing a campaign on changing patient behavior [2].

In marketing field has been studied on behavior and motivations of opinion leaders, which were summarized in the table 1 and divided into two groups of behaviors. The first group is covert behaviors (express behavior, but cannot be observed) and the second is overt behaviors (express behavior and can observed).

Table 1 Prior studies on opinion leader behaviors

Authors	Covert behaviors	Overt behaviors
Chan & Misra [4]	Public Individual, Personal Involvement	Product Familiarity, Print Media Exposure, Risk Preference
Weimann [5]	Personality Trait	Network Position
Flynn et al. [6]	Perceived Knowledge, Innovativeness	Enduring Involvement, Read, Shop, Spend Time
Liu [7]	Expertise	Capacity to store messages, Access News Media, Selective Perception, Interaction with social

Table 1 (continued) Prior studies on opinion leader behaviors

Authors	Covert behaviors	Overt behaviors
Sarathy & Patro [8]	Innovativeness	Product Knowledge, Media exposure, Social Involvement
Tsang & Zhou [9]	Extroversion	Enduring involvement, Personal-Product involvement

A tool for measuring the level of opinion leadership was first developed in form of questionnaire by Rogers and Cartano (1962) and improved by King and Summer (1970), which was a widely used self-assessment questionnaire. It was subsequently updated by Childer (1986) and considered aspect of psychometric properties by Flynn et al. [6]. Keller and Berry [10] used the criteria to identify opinion leaders at 10% of highest score. Watts and Dodds [11] argued that it remains unclear of criteria to consider influential persons, although some scholars used 32% of the highest score [12]. Moreover, some scholars have argued that self-assessment is one-sided information which may lead to bias. As a result, using information obtained by other approaches such as expert's assessment or voting by other members were also proposed. In the research of Coleman, Katz and Menzel [13], opinion leaders required voting form others at least 3 persons. Thus, our research has proposed a process for determining attributes and criteria to identify opinion leaders by using the decision tree algorithm based on the combined self-assessment and voting approaches.

2.2 Decision Tree Algorithm

The decision tree algorithm, which is a kind of supervised learning algorithm used for data classification was developed by J.R. Quinlan in 1986. It was used in a learning pattern of a training dataset and then created an equation or model for predicting from the new data. It has become a popular algorithm because the resulted model is easy to interpret and understand. The concept used to solve the problem was called "Recursive Partitioning" or "Divide and Conquer", which divided a complex problem into sub-problems until the sub-problems were easy enough to solve. After tackling each small subset, it will bring all answers together for solving the original problem [14]. There are currently many decision tree

algorithms that have been developed including ID3, C4.5, C5.0, ADTree, REP, FTTree, LADTree, Decision Stamp, LMT, CART and Random Forest [15].

One major advantage of the decision tree algorithm is that it provides a classification method that can support various types of data. It provides a learning process which can automatically create a model from the data, both numerical and categorical. It can select important features based on the data and can be applied to the cases where there are small amount of data for the system to learn. The results of the analysis model can be easily interpreted without complex mathematical formula, which makes it more popular than other models. One weakness of the decision tree is that the results are often inclined to majority group of the training dataset. Overfitting or underfitting may easily occur. In addition, small change to the training dataset can lead to a significant change of the result.

In this research, C4.5 algorithm was used for model learning as a baseline for comparing results with the Hellinger Distance Decision Tree algorithm (HDDT). The HDDT is an algorithm developed to accommodate imbalance data which may have problems with skewed class distribution or sparseness in feature space by changing the splitting criterion. By using the principle of Hellinger Distance on finding distributional divergence, these effects of the algorithm do not dominated by class priors.

3 Research methodology

This research consists of two sub-studies. The details are shown in Figure 1 and described as follows.

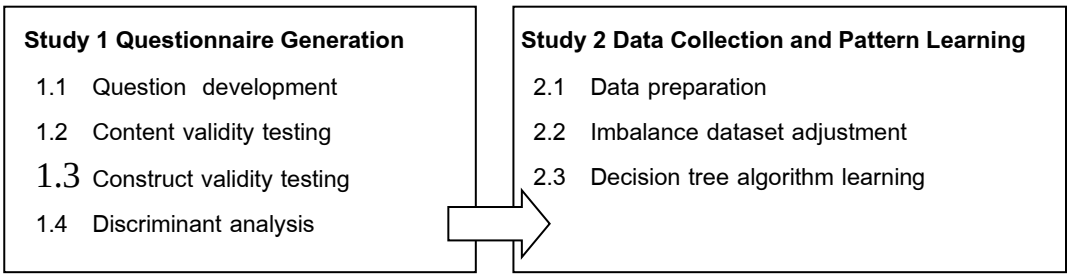

Figure 1 Overview of methodology

3.1 Study 1 Questionnaire Generation

The study consists of four steps as follows:

Step 1 Questionnaire development. The first step was to develop a questionnaire according to the cultural contexts of Thailand. A literature review was conducted for identifying important characteristics used to identify opinion leader. The questionnaire for measuring the opinion leadership levels by Flynn et al [6] was adapted for this research. Some independent variables include the questions for measuring about Knowledge, Word-of-Mouth, Self-confident and Extroversion. The full questionnaire consists of three parts: general information of the respondents, measurement of the level of opinion leadership and characteristics, and voting members of the group. In data collection, we developed a questionnaire for the domains of mobile phones and bicycles.

Step 2 Content validity testing. Testing validity of content consists of face validity on the questionnaire by ten people, which assessed the readability of the questions, and whether there was ambiguity. In addition, we conducted another test with five experts which examined the validity of content, then calculated the Index of Consistency (IOC) score and adjusted the questions based on the experts' comments.

Step 3 Construct validity testing. The structural validity was verified by using the Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA) techniques, which are parts of Structural Equation Modelling (SEM). The concept of CFA was used to test the psychometric properties of question in order to validate reliability and invariance of the factor structure when applied to different groups. The tests were conducted with 313 undergraduate students in Chiang Mai. The EFA test was conducted with 163 students and the CFA test was conducted with 150 students. Finally, the tests on the questionnaire confirmed a good reliability (Cronbach's alpha 0.913) resulting in 21 questions that include five factors. The results of CFA that compares the three models are shown in Table 2.

Table 2 Confirmatory Factor Analysis comparing result of the three models

Model	CMIN/DF	CFI	RMSEA	AIC
Null model	6.992	1.000	.201	1977.842
1-factor model	4.098	.528	.144	1128.753
5-factors model	1.738	.916	.070	415.612

Step 4: Discriminant Analysis testing. Finally, a test for the power of the questionnaire to distinguish different groups by testing on average score was conducted. First, the test score of 313 students were ranked. Then, the students were divided into two groups with a cutoff at the percentile of 25. Then, the independent sample t-test was used to test the difference of average score between the two groups. The result of t-test was found significantly different, meaning the questionnaire has a discriminant power.

3.2 Study 2 Data Collection and Pattern Learning

Step 1 Data preparation. Data were collected from a group of cyclists at Uttaradit province in Thailand. The collected self-assessment questionnaire responses of 120 members were used to generate the training and testing datasets. The training dataset was derived from answers of the questionnaires and the labels of opinion leaders are assigned for the persons who have the highest score of 4%, 6%, 10%, 16%, 20% and 25%. The reason that we used different cut-off scores was based on varying criteria from previous research. Thus, we can compare the results and choose the highest score level which give us the best prediction result. For the testing dataset, opinion leaders were labelled for the persons who received the votes more than 3 from other members in the group.

Step 2 Imbalance dataset adjustment. Due to the small number of opinion leaders comparing to the total number of members (minority group), the training dataset is found to have a problem of “Imbalance dataset”. Typical learning algorithms are usually appropriate for balance dataset. Thus, the learning result will give a very high accuracy on majority group, but very low on minority group. There are several approaches in solving the imbalance dataset problem but, the most commonly used techniques are Re-weighting, Over-sampling and Under-sampling [16]. However, Over-sampling may cause overfitting of the result and Under-sampling will lose some useful information. For this study, we followed suggestion form prior study on balancing data for C4.5 algorithm of Drummond and Holte [17], they had explained about the extra computational cost of Over-sampling is not always warranted for better performance. So both techniques were used, starting with Over-sampling and then followed by Under-sampling. In Weka, a data mining program developed by Machine Learning Group at the University of Waikato (https://www.cs.waikato.ac.nz/ml/weka/), package for Over-sampling called SMOTE (Synthetic Minority Over-sampling technique) and Spread Subsampling for Under-sampling are applied for generated additional balanced six

datasets. Other advanced techniques use the classifiers that are created especially for imbalance dataset such as Hellinger Distance Decision Trees (HDDT) [18]. Thus, we compared the prediction performance between four techniques: C4.5, Balanced+C4.5, HDDT, and Balanced+HDDT.

Step 3 Decision Tree Algorithm Learning with Weka was configuration to unprun. Evaluation performance of each model with testing dataset.

4 Results

The result in Table 3 showed the comparison of different Decision Tree learnings in Precision (P), Recall (R), and Accuracy (A) for class 1 (Opinion Leader group). In comparing the results from a study of the unbalance dataset on C4.5 and HDDT, it shows that HDDT give a higher overall accuracy than C4.5 in dataset 84, 80 และ 75. According to Cieslak et al. [18] mention that C4.5 are sensitive to the balance of data but HDDT are more stable. In the part of learning on balanced datasets, the result was not much different between C4.5 and HDDT, both of them give the highest overall accuracy at 81.25%

In deciding which model is the best, overall accuracy is not enough. The above mentioned that opinion leaders are a minority group of people in society, if we misclassify the actual opinion leader to non-opinion leader, we lose the valuable person for company. Such that, Recall is the measurement of interesting because it shows the correctly predicted observation over the total amount of actual class as “yes”. Whereas precision show correctly predicted positive observation to the total of predicted.

The best prediction result is obtained when the top-25% of the members was used as the cutoff score for opinion leaders and Balanced + C4.5 algorithm was applied to the dataset (P=83.33, R=50 and A=72.77). It is evident that the algorithm C4.5 can give good results if the data is more balanced. But, the HDDT is an algorithm developed based on the need to solve data imbalance, consequence to no sensitive in the balanced datasets However, HDDT performance is reduced in the more balanced datasets.

Table 3 Performance of classification model on opinion leader of different techniques

Percentile		**96**	**94**	**90**	**84**	**80**	**75***
Score		4%	6%	10%	16%	20%	25%
C4.5	P	0	0	0	100	100	66.67
	R	0	0	0	20	20	40
	A	50	50	50	50	50	40
Balanced + C4.5*	P	0	0	0	100	100	83.33
	R	0	0	0	20	20	50
	A	50	45	50	81.25	81.25	72.77
HDDT	P	0	0	0	100	100	66.67
	R	0	0	0	20	30	40
	A	50	50	50	81.25	80.30	67.77
Balanced + HDDT	P	0	0	100	100	100	66.67
	R	0	0	20	20	20	20
	A	50	45	81.25	81.25	81.25	75.30
Note: P= Precision, R = Recall, A = Overall Accuracy							

The important features extracted for the resulted decision tree model consist of Extroversion (EX), Knowledge (KN), and Self-confident (SF) as shown in Figure 2.

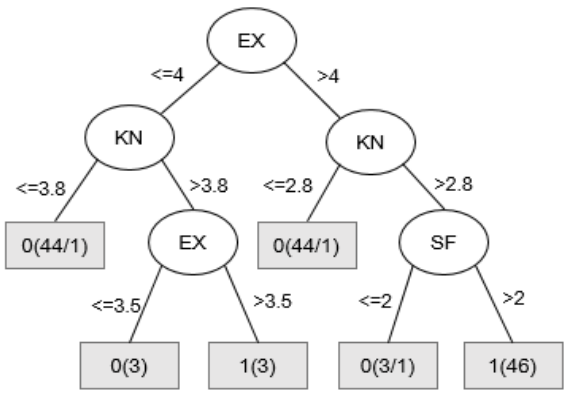

Figure 2 The resulted decision tree model

5 Conclusions

Identifying opinion leaders by learning from the self-assessment questionnaire and creating a prediction model by using the Decision Tree algorithm, showed the attributes used for prediction were Knowledge, Extroversion and Self-Confident. The opinion leaders are typically minority group in a society. This presents a problem known as the imbalance dataset problem when analyzing the data. Based on the evaluation results, the Balanced + C4.5 algorithm is found to result in better performance comparing to other techniques. Decision Tree of final model also revealed us how to develop the rule-based recommendation system.

According to the research of Chan & Misra, 1990 [4], they used Discriminant Analysis to classify opinion leader with resampling dataset. The ratio of opinion leader and non-opinion leader was 30:70 and report performance of prediction by precision, recall and accuracy as same as our study. For more insights to the result of learning model according to He & Ma, 2013 [19] recommend to report on curve-based metric such as ROC (Receiver Operating Characteristic) and AUC (Area under the curve) together with standardized assessment like Precision, Recall and Accuracy.

The limited for this research could be the amount of attribute used for prediction. Further study to better results can add more attribute such as innovativeness, mass media exposure or social network connection. The factors should be take into consideration for choosing attribute to study such as culture, domain of interesting and size of social groups. Finally, it is important to note that the key issues that affect the accuracy of forecasting are imbalance dataset which researcher should give more details on the method to balancing data and how to compare the result.

Acknowledgement

This research would not have been possible without the support of Faculty of Business Administration, Meajo University. Moreover the supported in part by Thailand Graduate Institute of Science and Technology (TGIST) National Science and Technology Development Agency (NSTDA) with contract number TGIST SCA-CO-2559-2300-TH and The National Research Council of Thailand.

References

[1] Katz E. The Two-Step Flow of Communication: An up-to-date report on an hypothesis. Public Opin Q 1957;21(1):61-78.

[2] Valente TW, Pumpuang P. Identifying opinion leaders to promote behavior change. Health Educ Behav 2006;34(6):881-96.

[3] Valente TW, Rogers EM. The origins and development of the diffusion of innovations paradigm as an example of scientific growth. Sci Commun 1995;16(3):242-73.

[4] Chan KK, Misra S. Characteristics of the opinion leader: a new dimension. J Advert 1990;19(3):53-60.

[5] Weimann G. Looking for opinion leaders: traditional vs. modern measures in traditional societies. Public Opinion Quarterly; 1991.

[6] Flynn LR, Goldsmith RE, Eastman JK. Opinion leaders and opinion seekers: two new measurement scales. J Acad Mark Sci 1996;24(2):137-47.

[7] Liu FCS. Constrained opinion leader influence in and electoral campaign season: revisiting the two-step flow theory with multi-agent simulation. Adv Complex Syst 2007;10(2):233-50.

[8] Sarathy PS, Patro SK. The role of opinion leaders in high-involvement purchased: an empirical investigation. South Asian J Manag 2013;20(2):127-145.

[9] Tsang ASL, Zhou N. Newsgroup participants as opinion leaders and seekers in online and offline communication environments. J Bus Res 2005;58(9):1186-93.

[10] Keller E, Berry J. The Influentials: one american in ten tells the other nine how to vote, where to eat, and what to buy. Soundview Executive Book Summaries; 2003.

[11] Watts DJ, Dodds PS. Influentials, networks, and public opinion formation. J Consum Res 2007;34(4):441-58.

[12] Coulter RA, Feick LF, Price LL. Changing faces: cosmetics opinion leadership among women in the new Hungary. Eur J Mark 2002;36(11/12):1287-308.

[13] Coleman J, Katz E, Menzel H. The diffusion of an innovation among physicians. Sociometry 1957;20(4):253.

[14] Lantz B. Machine learning with R: learn how to use R to apply powerful machine learning methods and gain an insight into real-world applications. Birmingham: Packt Publ; 2013.

[15] Sewaiwar P, Verma KK. Comparative study of various decision tree classification algorithm using WEKA. 2015 [cited 2016 May 1]; Available from: http://ermt.net/docs/papers/Volume_4/10_October2015/V4N10-113.pdf.

[16] Díez-Pastor JF, Rodríguez JJ, García-Osorio CI, Kuncheva LI. Diversity techniques improve the performance of the best imbalance learning ensembles. Inf Sci 2015;325: 98-117.

[17] Drummond C, Holte RC. C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. In: Workshop on learning from imbalanced datasets II. Citeseer; 2003. p. 1-8.

[18] Cieslak DA, Hoens TR, Chawla NV, Kegelmeyer WP. Hellinger distance decision trees are robust and skew-insensitive. Data Min Knowl Discov 2011;24(1):136-58.

[19] He H, Ma Y. Imbalanced learning: foundations, algorithms, and applications. John Wiley & Sons; 2013.

Author's Profile

Sudarat Sangkeaw, Ph.D. student, Faculty of Business Administration, Maejo University, Sansai, Chiangmai, 50290, sudarat_san@cmru.ac.th

Piyawan Siripraseotsin, Ph.D. Lecturer and Dean of Faculty of Business Administration, Maejo University, Sansai, Chiangmai, 50290, piyawan_nana@hotmail.com

Marut Buranarach, Ph.D. Senior Researcher at Language and Semantic Technology (LST) Laboratory, National Electronics and Computer Technology Center (NECTEC), 112 Thailand Science Park, Phahon Yothin Rd., Klong Luang, Pathumthani 12120, marut.bur@nectec.or.th