Investigation of Life Satisfaction Data by Data Mining and Machine Learning Techniques

ASLI KAYA
Department of Institutional Planning and Development
Eskisehir Technical University
Eskisehir
TURKEY

OZER OZDEMIR
Department of Statistics
Eskisehir Technical University
Eskisehir
TURKEY

FERDI KARAKUTUK
Zonguldak Regional Office
Turkish Statistical Institute
Zonguldak
TURKEY

Abstract: - With the exponential increase of data, which is called the mine of our age, from year to year, processing it and accessing information has gained more importance than ever before. The interest in data mining methods, which enable us to reach information from unprocessed data, is increasing in parallel with the increase in the amount of data. In this study, it is aimed to classify the 2019 Life Satisfaction Survey Hope Level variable data using data mining and machine learning algorithms and compare the algorithm results. As a result of the experimental studies, it has been seen that the k-NN algorithm makes more accurate and effective classification in Likert Scale data types.

Key-Words: - Classification, Data mining, k-NN algorithm, Support vector machine

1 Introduction

Analyzing this big data is a challenging process and hence the need for specific tools and techniques that are important in sorting large amounts of data becomes extremely important. Data Mining is one of the disciplines used to transform raw data into meaningful information and knowledge [1]. Data mining automatically searches and analyzes large volumes of data by discovering, learning and knowing hidden patterns, trends and structures [2] and answers questions that cannot be addressed with simple querying and reporting techniques [3]. Data Mining is generally divided into two categories [4], Predictive Data Mining: it can also be said that it is the model of the system that deals with the use of several attributes from a dataset and predicts future value or evolves according to the given data. On the other hand, Descriptive Data Mining: finds patterns that describe data, in other words, presents new insights based on current dataset trends.

Life satisfaction is a multidimensional concept shaped by various socio-demographic factors that lead to different expectations and preferences as well as different living conditions. Although women and men are almost equally satisfied, health status appears to be the main determinant of life satisfaction, ahead of factors such as financial situation, labor market situation or social relations.

The Life Satisfaction Survey, which has been carried out regularly since 2004, the first of which was implemented as an additional module in the Household Budget Survey in 2003; It aims to

measure the general happiness perception of the individual, social values, general satisfaction in basic living areas and satisfaction with public services and to follow the change of this satisfaction level over time [5].
In this study, it is aimed to classify the Life Satisfaction Research Hope Level Satisfaction data using data mining and machine learning algorithms and to compare the algorithm results.

2 Methods

2.1 Data Collection

In this study, the data of the Turkish Statistical Institute Life Satisfaction Survey were used. In this study, data were collected from 9212 individuals aged 18 and over in 4593 households by face-to-face interviews [6].

2.2 Data Preprocessing

The target variable to be classified within the scope of the study is B45- Hope level variable. While determining the hope level variable in the literature studies, it was not used for the target variable because of the indirect effects on this variable, where the aim of the questions other than the general satisfaction questions did not directly affect the hope level and general satisfaction. Data merging was performed on the variables B04, B05 and B06 in the data set. Updates made are shown in Table 1, 2 and Table 3.

Table 1: Working state data integration

Integrated Variable Code	Original Variable Values	Integrated Variable Values	Relationship Established Variables	Working_ variable
B04	Worked	Private	B04-B06	1
		Public		2
	Didn't work but continues to be related to his/her job	Didn't work but continues to be related to his/her job		3
	Didn't work	inability to find a job	B04-B05	4
		Working seasonally		5
B04	Didn't work	Continuing education/teaching	B04-B05	6
		busy with housework		7
		Retired		8
		Disabled or sick		9
		Old		10
		holder of will		11
		Other		12

Table 2: Status at work data editing

STATUS AT WORK	STATUS AT WORK (converted version)
11 paid or salaried	1
12 Daily wages (seasonal or casual jobs)	2
2 Employer	3
3 own account	4
4 unpaid family worker	5
not working	6

Table 3: Finished school data editing

SCHOOL_FINISHED	SCHOOL_FINISHED (converted version)
1 Didn't finish a school	1
2 Primary school	2
31 General secondary school	3
32 Vocational or technical secondary school	
33 Primary education	
41 General high school	4
42 Vocational or technical high school	
511 2 or 3 years college	5
512 4-year college or faculty	6
52 Master's (including 5 or 6-year faculties)	7
53 Doctorate	

2.3 Classification Algorithms

2.3.1 Decision Trees

A decision tree is a classifier expressed as a recursive division of the sample space. The decision tree consists of nodes that form a rooted tree, i.e. a directed tree with a node called a "root" without incoming edges. All other nodes have exactly one incoming edge. A node with outgoing edges is called an internal or test node. All other nodes are called leaves (also known as terminal or decision nodes). In a decision tree, each internal node divides the sample space into two or more subspaces according to a certain discrete function of the input attribute values. In the simplest and most common case, each test considers a single attribute such that the sample space is divided according to the value of the attribute. In the case of numeric attributes, the condition refers to a range.

Each leaf is assigned a class that represents the optimal target value. Alternatively, the leaf may hold a probability vector representing the probability that the target feature will have a given value. The specimens are classified by traversing from the root of the tree to a leaf according to the results of the tests carried out along the way.

In data mining, a decision tree is a predictive model that can be used to represent both classifiers and regression models. In operations research, decision trees refer to a hierarchical decision model and its results. The decision maker uses decision trees to determine the strategy most likely to achieve his goal.

2.3.1.1 J48 Algorithm

Classification is the process of constructing a class model from a set of records containing class labels. Decision Tree Algorithm is to find out how the feature vector behaves for a few samples. There are also classes for newly created examples based on training examples. This algorithm generates rules for the estimation of the target variable. With the help of tree classification algorithm, the critical distribution of the data can be easily understood. J48 is an extension of ID3. Additional features of J48 are accounting for missing values, pruning of decision trees, continuous attribute value ranges, derivation of rules. In WEKA data mining tool, J48 is an open source Java implementation of the C4.5 algorithms. The WEKA tool offers a number of options for tree pruning. In the case of potential overfitting, pruning can be used as a tool for refinement. In other algorithms, the classification is performed recursively until each leaf is pure, that is, the classification of the data should be as perfect as possible. This algorithm creates the rules by which the particular identity of this data is generated. The goal is to gradually generalize a decision tree until it gains a balance of flexibility and accuracy [7].

2.3.1.2 CART Algorithm

CART stands for Classification and Regression Trees [8]. It is characterized by forming binary trees, that is, each interior node has exactly two outgoing edges. Pods are selected using the bifurcation criterion and the resulting tree is pruned with cost-complexity pruning.

When provided, CART can take into account the costs of misclassification in tree induction. It also allows users to provide probabilistic distribution in advance. An important feature of CART is its ability to generate regression trees. Regression trees are trees whose leaves predict a real number, not a class. In the case of regression, the CART looks for bins that minimize the predicted squared error (least squares deviation). The prediction in each leaf is based on the weighted average for the node.

2.3.1.3 JRip Algorithm

This class implements a propositional rule learner. This algorithm was developed by William W. Cohen as an optimized version of IREP. Performs Repetitive Incremental Pruning to produce Error Reduction (RIPPER). JRip is a bottom-up method that learns rules by treating certain judgments of

examples in training data as a class and finding the set of rules that encompasses all members of the class. Cross-validation and minimum description length techniques are used to avoid overfitting [9].

2.4 Machine Learning

Machine learning is one of the fastest growing areas of computer science with wide-ranging applications. Machine learning is a subfield of computer science that developed from the study of pattern recognition and computational learning theory in artificial intelligence.
Machine learning explores the creation and study of algorithms that can learn from and make predictions on data [10]. Such algorithms work by constructing a model [11] from sample inputs to make data-based predictions or decisions, rather than strictly following static program instructions.
Machine learning is closely related to, and often overlaps with, computational statistics; A discipline that also specializes in forecasting. The field has strong ties to mathematical optimization, providing areas of methods, theory, and application.

2.4.1 k-NN Algorithm

The k-nearest neighbor (k-NN) algorithm is used to test the degree of similarity between the documents and k training data by determining the category of test documents and to store a certain amount of classification data. This method is a snapshot-based learning algorithm that categorizes objects according to the closest feature space in the training set. The training sets are mapped to the multidimensional feature space. The feature area is divided into regions according to the category of the training set. A point in the feature space is assigned to a particular category if it is the most frequent category among the k nearest training data. Usually Euclidean distance is typically used to calculate the distance between vectors. The key element of this method is the availability of a similarity measure to identify the neighbors of a particular document. The training phase consists of storing only the feature vectors and categories of the training set.
In the classification phase, the distances from the new vector representing an input document to all the stored vectors are calculated and the k closest samples are selected. The annotated category of a document is estimated based on the closest point assigned to a particular category [12].

2.4.2 Support Vector Machine Algorithm

The Support Vector Machine (SVM) algorithm is one of the discriminant classification methods that is widely considered to be more accurate. The SVM classification method is based on the principle of Structural Risk Minimization from computational learning theory [13]. The idea of this principle is to find a hypothesis that will guarantee the lowest true error. Moreover, SVM has solid foundations that are very open to theoretical understanding and analysis [14]. SVM needs both positive and negative training set, which is not common for other classification methods. This positive and negative training set is necessary for the SVM to search for the decision surface that best separates the positive from the negative in the n-dimensional space called the hyperplane. The document representatives closest to the decision surface are called support vectors. If documents that do not belong to support vectors are removed from the training dataset, the performance of the SVM classification remains unchanged [15].

3 Problem Solution

The aim of this study is to classify the hope level of individuals through life satisfaction and to compare data mining methods. Weka and Rstudio programs were used to implement decision tree and machine learning algorithms. The data set used in the application consists of 247 variables. The number of variables was reduced with expert opinion and algorithms were applied with 45 variables that were determined to play an important role in explaining the target variable. The k value was determined before applying the k-NN algorithm. Finding the value of k in k-NN is not easy. A small value of k means that noise will have a higher impact on the result and a large value will make it computationally expensive. Looking at the literature, there are many methods to determine the k-class value. In this study, the $k=\sqrt{n}$ formula was used to calculate the k value. However, before all algorithms are applied, the dataset is divided into 70% training data and 30% test data.
The results obtained in the algorithms are shown in Table 4.

Table 4: Comparative Survey Results

Algorithms	Kappa statistic	TP ratio	FP ratio	Certainty	F-criterion	ROC area	Accuracy	Number of rules	Duration (sec.)
J48	0,372	0,763	0,43	0,749	0,747	0,725	76,32%	143	0,11
JRip	0,355	0,756	0,43	0,741	0,74	0,672	75,61%	11	1,06
SimpleCart	0,355	0,758	0,44	0,744	0,741	0,695	75,85%	7	12,8
SVM	0,341	0,766	0,48	0,757	0,737	0,645	76,62%	-	20,94
k-NN	**0,856**	0,953	0,06	0,998	0,883	-	**94,57%**	-	0,11

The performance of the classification algorithm is usually examined by evaluating the accuracy of the classification. It is clearly seen in Table 4 that the k-NN algorithm has the highest classification accuracy (94.57%). It showed the second highest classification accuracy (76,62%) for the Support Vector Machines (SVM) algorithm, which is one of the important algorithms of Artificial Neural Networks. The JRip algorithm results in the lowest classification accuracy among the five algorithms (75.61%). For this reason, it has been seen that the k-NN algorithm outperforms other algorithms in terms of classification accuracy in Likert scale data types.

4 Conclusion

Life satisfaction means the level of positive feeling that individuals reach depending on their evaluation of their personal life quality as a whole. In recent years in our country, there has been a significant increase in studies on the level of happiness and hope in parallel with the trend in the world. In these studies, "how hopeful are you when you think about your own future?" questions are asked and hope levels are measured. The level of response varies from three scales to ten scales. In this research, we performed experiments to determine the classification accuracy of five algorithms that we think will better predict the hope level of individuals, with the help of the Rstudio program, with an attractive data mining tool known as WEKA.

As a result of the experiments, the classification accuracy of the k-NN algorithm was the highest in the data types using the Likert Scale. In the literature researches, it has been seen that the k-NN algorithm is not widely applied for Likert Scale data types. We think that our study will contribute to the literature in this regard.

References:

[1] Rokach L., Maimon O. (2008). "Data mining with decision trees". Theory and applications, SERIES IN MACHINE PERCEPTION AND ARTIFICIAL INTELLIGENCE, Volume 69.

[2] Aggarwal, C. C. (2015). "Data Mining: The Textbook". Springer, 2015.

[3] El-Deen Ahmeda, R. A., Shehaba, M. E., Morsya S. and Mekawiea, N. (2015). "Performance Study of Classification Algorithms for Consumer Online Shopping Attitudes and Behaviour Using Data Mining. In Communication Systems and Network Technologies (CSNT)", Fifth International Conference on IEEE, pp. 1344-1349.

[4] Gnanapriya, S., Suganya, R., Devi, G. S. and Kumar, M. S. (2010). "Data Mining Concepts and Techniques". Data Mining and Knowledge Engineering, vol. 2, p. 256-263.

[5] Türkiye İstatistik Kurumu, Yaşam Memnuniyeti Araştırması Mikro Veri Seti 2020, ISBN 978-605-7613-71-4.

[6] Türkiye İstatistik Kurumu https://www.tuik.gov.tr/Kurumsal/Mikro_Veri.

[7] Kaur, G., & Chhabra, A. (2014). "Improved J48 classification algorithm for the prediction of diabetes". International Journal of Computer Applications, 98(22).

[8] Breiman L., Friedman J., Olshen R., and Stone C. (1984) Classification and RegressionTrees. Wadsworth Int. Group.

[9] Veeralakshmi V., Ramyachitra, D. (2015). "Ripple Down Rule learner (RIDOR) Classifier for IRIS Dataset". Issues, vol 1, p. 79-85.

[10] Kohavi, R., Provost, F. (1998). “Glossary of terms”. Machine Learning 30: 271–274.
[11] Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer. ISBN 0-387-31073-8.
[12] Tam, V. H., Santoso, A., Setiono, R. (2002). A comparative study of centroid-based, neighborhood-based, and statistical approaches for effective document categorization. In: Object recognition supported by user interaction for service robots. IEEE, p. 235-238.
[13] Cortes, C., Vapnik, V. (1995). “Support-vector networks”. Machine Learning 20 (3): 273. doi:10.1007/BF00994018.
[14] Press, W. H., Teukolsky, S. A., Vetterling, W. T., Flannery, B. P. (2007). “Section 16.5. Support Vector Machines”. Numerical Recipes: The Art of Scientific Computing (3rd ed.). New York: Cambridge University Press. ISBN 978-0- 521-88068-8.
[15] Ferris, M. C., Munson, T. S. (2002). “InteriorPoint Methods for Massive Support Vector Machines”. SIAM Journal on Optimization 13 (3): 783. doi:10.1137/S1052623400374379.