Mühendislik Bilimleri ve Tasarım Dergisi
11(4), 1257-1269, 2023
e-ISSN: 1308-6693

Journal of Engineering Sciences and Design
DOI: 10.21923/jesd.1121792

Araştırma Makalesi Research Article

A NEW RATIONAL CLASSIFICATION APPROACH BY THE NEW MIXED DATA BINARIZATION METHOD

Muhammed SÜTÇÜ[1,2*], İbrahim Tümay GÜLBAHAR[1]

[1]Abdullah Gül University, Faculty of Engineering, Department of Industrial Engineering, Kayseri, Türkiye
[2]Engineering Management Department, College of Engineering & Architecture, Gulf University for Science & Technology, Mishref 32093, Kuwait

Keywords

Binarization, Classification, Data Mining, IRIS Data Set, Decision Models.

Abstract

Classification algorithm is a supervised learning technique that is used to identify the category of new observations. However, in some cases, quantitative and qualitative data must be used together. With this approach, we tried to overcome the problems encountered in using quantitative and qualitative data together. In this paper, we model a new classification technique by converting all types of data to binary data because in the real world, data are classified in different types such as binary, numeric, or categorical. By this way, we develop a more accurate and efficient mixed data binarization approach for multi-attribute data classification problems. First, we determine the classes from available dataset and then we classify the new instances into these predetermined classes by using the new proposed data binarization approach. We show how each step of this algorithm could be performed efficiently with a numeric example. Then, we apply the proposed approach on a well-known iris dataset and our model show promising results and improvements over previous approaches.

KARMA VERİ İKİLİLEŞTİRME YÖNTEMİ İLE YENİ BİR RASYONEL SINIFLANDIRMA YAKLAŞIMI

Anahtar Kelimeler

İkilileştirme, Sınıflandırma, Veri Madenciliği, IRIS Veri Seti, Karar Modelleri.

Öz

Sınıflandırma algoritması, yeni gözlemlerin kategorisini belirlemek için kullanılan denetimli bir öğrenme tekniğidir. Ancak bazı durumlarda nicel ve nitel verilerin birlikte kullanılması gerekir. Bu yaklaşımla nicel ve nitel verilerin birlikte kullanılmasında karşılaşılan sorunlar aşılmaya çalışılmıştır. Bu çalışmada, gerçek dünyada veriler ikili, sayısal veya kategorik gibi farklı türlerde sınıflandırıldığından, tüm veri türlerini ikili verilere dönüştürerek yeni bir sınıflandırma tekniği modellenmektedir. Bu sayede çok özellikli veri sınıflandırma problemleri için daha doğru ve verimli bir karma veri ikilileştirme yaklaşımı geliştirilmiştir. Öncelikle mevcut veri setinden sınıfları belirlenmektedir ve ardından yeni önerilen veri ikilileştirme yaklaşımını kullanarak yeni örnekleri bu önceden belirlenmiş sınıflara sınıflandırılmaktadır. Bu algoritmanın her adımının nasıl verimli bir şekilde gerçekleştirilebileceğini sayısal bir örnekle gösterilmiştir. Ardından, önerilen yaklaşımı iyi bilinen bir iris veri kümesine uygulamış ve modelimiz önceki yaklaşımlara göre umut verici sonuçlar ve iyileştirmeler verdiği gösterilmiştir.

Alıntı / Cite

Sutcu, M., Gulbahar, I.T., (2023). A New Rational Classification Approach by The New Mixed Data Binarization Method, Journal of Engineering Sciences and Design, 11(4), 1257-1269.

Yazar Kimliği / Author ID (ORCID Number)		Makale Süreci / Article Process	
M. Sutcu,	0000-0002-8523-9103	**Başvuru Tarihi / Submission Date**	26.05.2022
I.T. Gulbahar,	0000-0001-9192-0782	**Revizyon Tarihi / Revision Date**	12.07.2023
		Kabul Tarihi / Accepted Date	15.08.2023
		Yayım Tarihi / Published Date	30.12.2023

[*]İlgili yazar/Corresponding author: muhammed.sutcu@agu.edu.tr, +90-352-224-8800

A NEW RATIONAL CLASSIFICATION APPROACH BY THE NEW MIXED DATA BINARIZATION METHOD

Muhammed SUTCU[1,2*], Ibrahim Tumay GULBAHAR[1]
[1]Abdullah Gül University, Faculty of Engineering, Department of Industrial Engineering, Kayseri, Türkiye
[2]Engineering Management Department, College of Engineering & Architecture, Gulf University for Science & Technology, Mishref 32093, Kuwait

Highlights

- Introduction of a novel mixed data binarization technique for multi-attribute data classification.
- Successful integration of diverse data types, including binary, numeric, and categorical.
- Improved classification accuracy and efficiency demonstrated through practical applications.
- Valuable contribution to addressing challenges in mixed data classification and enhancing data analysis.

Purpose and Scope
In order to increase the accuracy of multi-attribute data classification, this research offers a ground-breaking classification technique designed to successfully integrate various data types. The study deals with problems arising from binary, quantitative, and categorical real-world datasets with various data classifications. The goal is to create a reliable process that converts various data types into binary representations, making it possible to classify new instances accurately and quickly.

Design/methodology/approach
The research suggests a comprehensive mixed data binarization method. This method addresses computational complexity and classification accuracy difficulties by converting various data sources into binary representations. The program recognizes the dataset's data classes before using the cutting-edge mixed data binarization method to categorize fresh instances. This approach is demonstrated with a detailed numerical example and its implementation on a well-known iris dataset.

Findings
The suggested mixed data binarization technique achieves encouraging results, boosting classification accuracy greatly and addressing issues brought on by various data kinds. The method offers an effective approach for multi-attribute data classification by successfully merging quantitative and qualitative data. The results of the study highlight the significance of precise class determination within the dataset as a prerequisite for the proper categorization of new occurrences.

Originality
This research presents a revolutionary methodology that advances the field of multi-attribute data classification by seamlessly integrating several data types. The proposed mixed data binarization method offers reusability, increased precision, and computational effectiveness. This novel method is useful for academics, practitioners, and educators looking for precise data classification techniques because it may be used in a variety of situations.

* İlgili yazar/Corresponding author: muhammed.sutcu@agu.edu.tr, +90-352-224-8800

1. Introduction

Data classification is also very popular with the advances in technology which helps to increase capability of both generating and collecting data. This leads to a trend for data, its size and dimensionality grow. The widespread use of labeling techniques like barcodes, the computerization of businesses and advances in the data collection tools have provided us a huge amount of data. Millions of databases are now used by companies, governments, and universities (UCI, 2007; University of Toronto, 2003). It is noted that the number of these databases continue to grow rapidly because of the high-tech database systems. So, mining the unrefined data and convert it into useful information and knowledge is very important. However, there are a lot of different attributes related to stored data. Some of them have small some others have big impacts on decisions. Thus, splitting up the relevant and irrelevant data becomes an important issue.

Separation of relevant and irrelevant attributes becomes important because irrelevant attributes contain little or no information, for example, students' ID is often irrelevant to the task of predicting students' GPA. Also "redundant attributes" duplicate much, or all of the information contained in one or more other attributes, for example, purchase price of a product and the amount of sales tax paid. Moreover, we should create new attributes that can capture the important information on a data set which are much more efficiently than the original attributes.

A large number of data classification methods have been developed, but they have some obstacles and difficulties which make them unattractive (Carnevalli and Miguel, 2008). Hence, researchers are focusing on developing more accurate and more efficient methods or improving the existing methods. Some of the previous methodologies have computational difficulties for large data sets, also these methodologies are time consuming and too expensive. Also, previous approaches are limited to either continuous or discrete case. Moreover, previous approaches are not enough to handle them, and they don't predict efficiently the classes if the dataset includes different variable types. In order to overcome these difficulties, we construct a new classification technique by converting all types of data to binary data because in real world, data are classified in different types such as binary, numeric or categorical. This issue leads us to develop a more accurate and efficient mixed data binarization approach for multi-attribute data classification problems.

Increasingly, methodologists have emphasized the integration of qualitative and quantitative data as the centerpiece of mixed methods. Integration is an intentional process by which the researcher brings quantitative and qualitative approaches together in a study. There are bunch of machine learning tools used in the literature, however they do not give good results due to some limitations. Most of the research data set integrates qualitative and quantitative data in a single research study. It involves collecting and analyzing qualitative and quantitative data to understand a phenomenon better and answer the research questions. Researchers combine qualitative and quantitative methods and data to expand their evidence, improve the credibility of their findings, and illustrate the results from one method with the results from the other one.

In this study, we propose a tool and a model that brings together various quantitative and qualitative data analysis (i.e., mixed analysis) techniques into one meta-framework to assist researchers who use qualitative and quantitative approaches within the same study in the data analysis phase. Our approach combines different types of data by converting them into binary representations. By doing so, we establish a more accurate and efficient mixed data binarization technique for multi-attribute data classification problems. In this paper, we begin by identifying various data types commonly encountered in real-world scenarios, such as binary, numeric, and categorical data. Our proposed approach focuses on converting these diverse data types into binary format to facilitate effective classification. We emphasize the importance of accurately determining the classes within the available dataset, which serves as a foundation for subsequent classification of new instances.

By presenting this novel technique, we aim to contribute to the field of classification algorithms by addressing the challenges associated with integrating diverse data types. Our approach offers a practical solution that can be applied to a wide range of real-world scenarios. The findings of this study provide valuable insights and open new avenues for future research in the field of data classification.

The remainder of this paper is structured as follows: In section 2, we present methodology and theory of the new classification approach. Section 3 discusses the results of the application of the new classification approach to buying computer example. Section 4 contains concluding remarks.

2. Literature Review

Data Classification is one of the most important working areas in data mining. Data classification is a supervised

learning strategy that categories the data in distinct classes. It deals to identify the patterns and classify the new samples into known classes (Silva and Zhao, 2012; Schwenker and Trentin 2014). Classification problems have been studied by different researchers including computer scientists, engineers, statisticians, biologist, and economists (Jouni et al., 2014; Buniyamin et al., 2016; Loh, 2011; Graur et al., 2015).

There are variety of methods for solving classification problems such as neural networks, K-nearest neighbor approach, support vector machines, linear programming, and fuzzy logic (Pratikakis et al., 2017; Cover and Hart, 1967; Zhang et al., 2018; Pal and Foody, 2010; Bai, 2020; Melin et al., 2013). Many data classification methods have been developed till now, but each of them has some shortcomings and difficulties which make them unattractive. Therefore, researchers are focusing on developing more accurate and efficient methods or trying to improve the existing methods. This leads different classification models to be applied in different fields in the literature including finance, risk management, health care, sports, engineering, and science (Zhang et al., 2013; Singhal et al., 2011; Faes, 2019; Russo et al., 2019; Waltman and van Eck, 2012). Also, data classification has a wide range of customer segmentation related applications as well. Characterization of customer segmentation into groups with similar behaviour and predict customer purchasing behavior such as buying a car, can be identified by classification (Sutcu, 2020).

Classification approaches are widely used in various fields to categorize and allocate labels to different entities or data points. In the field of supplier evaluation and selection, Ho et al. (2010) conducted a literature review to survey the multi-criteria supplier evaluation and selection approaches. They classified and analyzed international journal articles from 2000 to 2008, extending previous reviews that covered literature up to 2000. This review contributes to the understanding of supplier evaluation and selection models and provides insights into the classification of approaches in this domain. In the context of hyperspectral image classification, Ghamisi et al. (2017) reviewed existing spectral classifiers specifically developed for hyperspectral images. They highlighted the challenges in hyperspectral image classification, such as the presence of redundant features, limited training samples, and high dimensionality of the data. The review compared different classification approaches, including support vector machines, random forests, neural networks, deep approaches, and logistic regression-based techniques, in terms of classification accuracy and computational complexity. Their review provides a comprehensive overview of spectral classification approaches for hyperspectral images.

Classification-based approaches for spam detection in social networks face limitations in data labeling and spam drifting (Koggalahewa et al., 2021). (Koggalahewa et al., 2021) highlighted the difficulty of detecting new forms of attacks using classifiers trained on older datasets. This review emphasizes the challenges and limitations of classification approaches for spam detection in social networks. L'Hermitte et al. (2014) proposed a new classification model for disasters based on their logistics implications. They conducted a literature review to develop a theory-based approach and applied the model to a case study of the 2011-2012 Somali food crisis. The review integrated situational factors that reflect the impact of the external environment on logistics operations. This study contributes to the understanding of disaster classification from a logistics perspective. In the field of Autism Spectrum Disorder (ASD), Wolfers et al. (2019) reviewed pattern classification and stratification approaches used in research on ASD. They observed large variance in predictive performance across pattern classification studies and discussed factors contributing to this variance, such as sampling bias, validation procedures, and data quality. This review highlights the challenges in mapping biological differences and individual trajectories in ASD. The TNM classification system for malignant tumors has undergone significant revisions over the years (Webber et al., 2014). (Webber et al., 2014) described the annual literature review process used to inform the revisions of the TNM classification.

Expert panels reviewed scientific literature on staging and provided feedback for revisions. This review provides insights into the evidence-based process of improving the TNM classification. Early classification of time series has been extensively studied for time-sensitive applications (Gupta, 2020). (Gupta, 2020) conducted a systematic review of existing approaches for early classification of time series. The review categorized the approaches into four exclusive categories based on their solution strategies. This review provides a comprehensive overview of early classification approaches for both univariate and multivariate time series. Varese & Lombardi (2020) conducted a literature review to analyze the concept, classification, functionalities, and technological processes of dry ports. They used a literature review as a valid approach to identify gaps, issues, and opportunities for further study and research in the field of dry ports. Traffic classification plays a crucial role in prioritizing, protecting, and preventing certain types of Internet traffic (Dainotti et al., 2012). (Dainotti et al., 2012) discussed the challenges and recent advances in traffic classification capabilities. They emphasized the importance of situational awareness of traffic for preventing and mitigating new forms of malware. This review highlights the need for research on global Internet traffic characteristics and the implications of traffic classification for network optimization. Stakeholder identification and classification are essential in sustainability marketing (Kumar et al., 2016). (Kumar et al., 2016) reviewed the literature on stakeholder identification and classification in sustainability marketing

from 1998 to 2012. They provided a generalized approach to stakeholder identification and classification, contributing to the understanding of stakeholder management in the context of sustainability marketing. approaches used in these fields and contribute to the understanding and advancement of classification methodologies.

The objective in data classification is to assign instances that are described by several attributes into a predefined number of classes. The nearest neighbor technique firstly proposed by Fix and Hodges (1951), then modified by Cover and Hart (1967). The use of K-nearest Neighbor for defining the unlabeled samples based on their similarity with the observations in the training set. Nevertheless, experimental studies show that K-nearest neighbor is computationally expensive for a large data set. To find the optimum class, the approach runs several times until the sample converges to a class. Moreover, it needs a large storage, because it runs using the whole set and highly sensitive to the curse of dimensionality. Moreover, it is not possible to quickly reject candidates by using the difference in one coordinate as a lower bound for a distance based on all the dimensions. On the other hand, K-nearest neighbor is simple and faster than other classification methods. Moreover, the misclassification rate of k-NN rule approaches the optimal error rate asymptotically as "k" increases.

3. Material and Method

In this section, we first discuss the new proposed approach and then explain and motivate the new approach with synthetic data. We then describe the matrix types used in our new method and we will give their mathematical formulations. We finally explain applicability of the proposed model with a motivating example.

3.1. New Data Classification Approach

In this subsection, we discuss the theory of the binarization approach.

3.1.1. Theory of the Binarization Approach

The data classification problem is considered in two parts as training part and testing part. Determination of the characteristics of the instances that belong to a certain class and differentiating them from the instances that belong to other classes are the main objectives of the training part. After the classes are determined, then by using the testing part, we can find the classes of the new instances.

Given the classes, the attributes and values of the attributes are sufficient to solve the classification problem with binarization approach. The all data set is represented by set Z. We divide the dataset Z into two groups; training data set which is represented by X, and test data set which is represented by Y. Xi where i=1, …, n and Yi where j=1, …, m show the data in each set. There is totally (n+m) amount of data in the set Z. Then, the attributes are shown by

$$Values\ of\ instance\ X_i = \{x_{i1}, x_{i2}, \dots, x_{ik}\} \quad (1)$$

$$Values\ of\ instance\ Y_j = \{y_{j1}, y_{j2}, \dots, y_{jk}\} \quad (2)$$

where k=number of attributes, x_{ik}, y_{jk} are the attribute values of kth attribute of ith and jth data. We have k attributes, and each attribute has at least one value. For instance, we have 3 attributes A, B, and C. A and B have 2 attribute values, C has 3 attribute values. So, we can represent them as $A = \{a_1, a_2\}$, $B = \{b_1, b_2\}$, and $C = \{c_1, c_2, c_3\}$.

For a given dataset, each data in dataset is written as binary values. So, "1" or "0" is used as an attribute value instead of using the real values. Assume instance-1 has the attribute values like $X_1 = \{a_1, b_2, c_3\}$ so, this attribute values are converted to a binary case by $A = [1,0]$, $B = [0,1]$, $C = [0,0,1]$ and then in the same logic, the binary vector is written as $X_1^b = [1,0,0,1,0,0,1]$ where X_1^b is the binary vector of instance-1.

3.1.2. Classification of Unclassified Instances

For determining the classes of unclassified instances, we use the metric distance approach. Generally, the aim is to minimize the distance between two classified instances, while maximize the distance between different classes. The most commonly used metrics to measure the distance of a sample from a given training set $X = \{x_1, x_2, \dots, x_n\}$ are as follows where formulation 3 is Minkowski distance, formulation 4 is Least Square distance, formulation 5 is

Manhattan distance, and formulation 6 is Chebychev distance:

$$d(X, X^*) = \left(\sum_{i=1}^{n} |x_i - x_i^*|^r \right)^{\frac{1}{r}} \quad (3)$$

$$d(X, X^*) = \sqrt{\sum_{i=1}^{n} (x_i - x_i^*)^2} \quad (4)$$

$$d(X, X^*) = \left(\sum_{i=1}^{n} |x_i - x_i^*| \right) \quad (5)$$

$$d(X, X^*) = max_i^n |x_i - x_i^*| \quad (6)$$

where X* is the unclassified instance data set, x_i^* is the i[th] unclassified instance. Also, here "d" represents the distance measure, "r" represents the order parameter and finally "n" is the number of instances.

In this study, the Least Square Distance approach is used to measure the performance of the proposed model. The distance between unclassified instance and each class is calculated to find the nearest class for the instance. For each class, $d(X_j, X^*) = (\sum_{i=1}(x_{ji} - x_i^*)^2)$ is used as distance measure where X* is the unclassified data, j is jth class and i is the ith value of the class j and the value of unclassified data. After explaining the previous distance methods, we now introduce our new measurement approach which is called multiplication approach by using a synthetic data set.

3.2. Example of the New Binarization Model with Synthetic Data

We use a synthetic dataset to explain and motivate the new approach. The data table is shown in Table 1. There are four attributes, of which X and W are integer, Y is continuous, and Z is categorical. There are 5 classes in this example.

Table 1. Motivating Example Data Set

INITIAL MIXED DATA SET

X	Y	Z	W	
3	0.342	A	3	X=Integer
5	0.51	C	120	Y=Continuous
2	0.104	E	141	Z= Categorical
1	0.779	B	6	W=Integer
4	0.628	D	130	

Firstly, attribute "W" is converted to nominal data by clustering with k-Nearest Neighbor algorithm. In this algorithm, the centroid of a cluster is selected as its center point. The centroid can be defined in various ways such as by the mean or medoid of the objects (or points) assigned to the cluster. The difference between an object $p \in C_i$ and c_i , the representative of the cluster, is measured by $dist(p, c_i)$, where $dist(x, y)$ is the Least Square distance between two points x and y. The quality of cluster C_i can be measured by the within cluster variation, which is the sum of squared error between all objects in C_i and the centroid c_i, defined as

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} dist(p, c_i)^2 \quad (7)$$

where E is the sum of the squared error for all objects in the data set; p is the point in space representing a given

object; and c_i is the centroid of cluster C_i (both p and c_i are multi-dimensional). Here we have the values $\{3, 120, 141, 6, 130\}$ for the attribute W. Intuitively, by visual inspection we may imagine the points partitioned into the two clusters, if equation-7 is applied and if we select k as 2, the partitioning $\{3, 6\}$ and $\{120, 141, 130\}$ has the within cluster variation as following:

$$(3 - 4{,}5)^2 + (6 - 4{,}5)^2 + (120 - 130{,}3)^2 + (130 - 130{,}3)^2 + (141 - 130{,}3)^2 = 238{,}37 \qquad (8)$$

given that the mean of cluster $\{3, 6\}$ is 4.5 and the mean of $\{120, 141, 130\}$ is 130.3. Therefore, 238.37 is the lowest distance between classes when we partition the set into 2 clusters. Then, the set is converted into numeric values as $w_1 = \{3, 6\}$ and $w_2 = \{120, 141, 130\}$. So, the new dataset converts to nominal values by k-NN and now an updated data set with the attribute values w1 and w2 of attribute W is generated. Table-II shows the values of the updated version of the dataset.

Table 2. Motivating Example Converted Data Set

X	Y	Z	W
3	0,342	A	W_1
5	0,51	C	W_2
2	0,104	E	W_2
1	0,779	B	W_1
4	0,628	D	W_2

The second step is converting the continuous attributes by splitting method as shown in Table-3. If the training dataset is big, we could assume that these values are distributed normally. So, we can find the mean and the variance of an attribute and then convert each value of an attribute to a probability value. Then we use the splitting method and partition the continuous valued attributes into 2, 3 or more branches. For instance, if we partition into two, then two branches are grown, corresponding to $A \leq split_{point}$ and $A > split_{point}$. Then, the data values are converted to binary values. (Assume $split_{point}$ = 0,50)

Table 3. The Binary Data Set for New Approach

X					Y		Z					W	
X_1	X_2	X_3	X_4	X_5	Y_1	Y_2	Z_1	Z_2	Z_3	Z_4	Z_5	W_1	W_2
0	0	0	1	0	1	0	1	0	0	0	0	1	0
0	0	1	0	0	0	1	0	0	0	1	0	0	1
1	0	0	0	0	1	0	0	1	0	0	0	0	1
0	0	0	0	1	0	1	0	0	0	0	1	1	0
0	1	0	0	0	0	1	0	0	1	0	0	0	1

The final step is the finding the most suitable class for the unclassified instance. For this, Least Square Distance method is used to solve the minimization problem. The class is found by

$$Minimum\ distance = \min_j \left(d(X_j, X^*) = \left(\sum_{i=1} \left(x_{ji} - x_i^* \right)^2 \right) \right) \qquad (9)$$

3.3. Class Matrix, Preference Matrix and Decision Matrix

In this section, we will describe the matrix types used in our new method and we will give their mathematical formulations.

Class matrix shows each class and their attribute values. It is a matrix where rows show the classes and columns show the attribute values.

$$Class\_Matrix\ C_{(m\times n)} = \begin{pmatrix} c_{11} & \cdots & c_{1n} \\ \vdots & \ddots & \vdots \\ c_{m1} & \cdots & c_{mn} \end{pmatrix} \quad (10)$$

where m=number of classes, n=total number of values of attributes.

Preference matrix shows the preferences of decision maker as a column vector.

$$Preference\_MatrixP_{(n\times 1)} = \begin{pmatrix} p_{11} \\ p_{21} \\ \cdots \\ \cdots \\ p_{n1} \end{pmatrix} where\ n = \left(\sum_{i=1}^{x} v_i\right) \quad (11)$$

where v_i is the number of values of an attribute.

Finally, the decision matrix is the multiplication of the class matrix and preference matrix which shows the results of the distances between each class and unclassified distance. We can easily read the distances from the decision matrix. We can find the most suitable class by reading the maximum value of each value of the matrix.

$$DecisionMatrix\ D_{(m\times 1)} = C_{(m\times n)} \times P_{(n\times 1)} = \begin{pmatrix} c_{11} & \cdots & c_{1n} \\ \vdots & \ddots & \vdots \\ c_{m1} & \cdots & c_{mn} \end{pmatrix} \times \begin{pmatrix} p_{11} \\ p_{21} \\ \cdots \\ \cdots \\ p_{n1} \end{pmatrix} = \begin{pmatrix} d_{11} \\ d_{21} \\ \cdots \\ \cdots \\ d_{n1} \end{pmatrix} \quad (12)$$

We then look at the best class for a customer's preferences by $c^* = argmax_i = (d_{i1})$ where c* shows the best suitable class.

3.4. The Motivating Example of the New Classification Approach

The motivating example includes 4 attributes where attributes "age" and "income" have three values, attributes "student" and "credit" have 2 values. So, for this motivating example, there are different classes however handling with these all classes is difficult, expensive and time consuming. Because this is a small example, but it has 36 different classes. As handling with a bigger problem with so many attributes and values, the number of classes and difficulty of the problem increases exponentially.

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

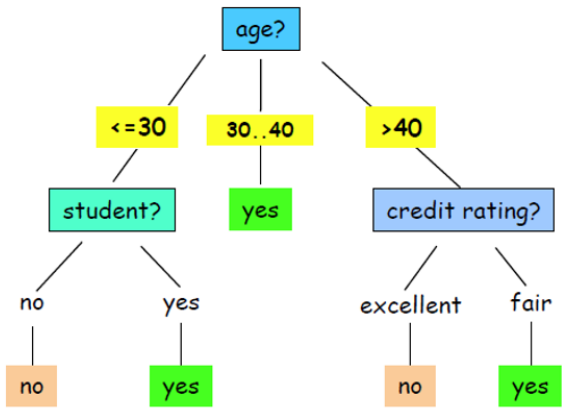


Figure 1. Classification of the Motivation Example by Decision Tree Induction

Due to the difficulty of the problem, we prune the number of classes to fourteen instead of seventy-two. The attributes are given as

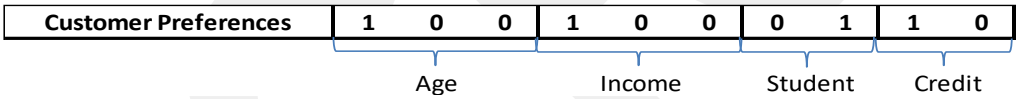
$$Age = \{under\ 30, 31-40, above\ 40\}$$
$$Income = \{low, medium, high\}$$
$$Student = \{no, yes\} \quad (13)$$
$$Credit\_rating = \{fair, excellent\}$$

Next step, we convert the values of each class into binary values. Therefore, the new table is shown in Table 4.

Table 4. The Binary Data Set for Buying Computer Case

	AGE			INCOME			STUDENT		CREDIT		OUTCOME	
	Age-T	Age-M	Age-O	Income-L	Income-M	Income-H	Student-N	Student-Y	Credit-F	Credit-E	Buy	Not buy
Class 1	1	0	0	0	0	1	1	0	1	0	0	1
Class 2	1	0	0	0	0	1	1	0	0	1	0	1
Class 3	0	1	0	0	0	1	1	0	1	0	1	0
Class 4	0	0	1	0	1	0	1	0	1	0	1	0
Class 5	0	0	1	1	0	0	0	1	1	0	1	0
Class 6	0	0	1	1	0	0	0	1	0	1	0	1
Class 7	0	1	0	1	0	0	0	1	0	1	1	0
Class 8	1	0	0	0	1	0	1	0	1	0	0	1
Class 9	1	0	0	1	0	0	0	1	1	0	1	0
Class 10	0	0	1	0	1	0	0	1	1	0	1	0
Class 11	1	0	0	0	1	0	0	1	0	1	1	0
Class 12	0	1	0	0	1	0	1	0	0	1	1	0
Class 13	0	0	0	0	0	1	0	1	1	0	1	0
Class 14	0	0	1	0	1	0	1	0	0	1	0	1

Now, our model is ready for the unclassified instances. For instance, one of the unclassified instances is given a teen-age, low-income student with a fair credit score. So, it can be written $X = \{teen, low, yes, fair\}$. Then we convert it to binary values by



Finally, we calculate the distances between the instance and 14 classes one by one. The results are shown in Table 5. Also, the outcomes of each class are shown in Table 6.

Table 5. Least Square Distances of Unclassified Instance

Class 1	6	Class 8	6
Class 2	8	Class 9	2
Class 3	4	Class 10	4
Class 4	6	Class 11	6
Class 5	2	Class 12	6
Class 6	4	Class 13	3
Class 7	2	Class 14	8

Table 6. Outcomes of each Class

Class 1	not buy	Class 8	not buy
Class 2	not buy	Class 9	Buy
Class 3	Buy	Class 10	Buy
Class 4	Buy	Class 11	Buy
Class 5	Buy	Class 12	Buy
Class 6	not buy	Class 13	Buy
Class 7	Buy	Class 14	not buy

We can see from the example, without spending too much time and money, we can easily find the class of the unclassified instance. Here, we can see that the instance would be in class-5, class-7 or class-9. In all the options,

we can say that the customer is going to buy a computer. Also, if we compare our findings with the test data (the unclassified instance was taken from the test data set), the customer's behavior is to buy a computer. Also, we try several options and for each case our findings are 100% correct.

4. An Application on IRIS Flower Dataset

The IRIS data set were used by Fisher in his development of the linear discriminant function and is still one of the standard discriminant analysis examples used in explaining or testing most current approaches and methodologies (Fisher, 1936). In this problem, three classes of IRIS flowers are to be discriminated using four continuous valued features that represent physical characteristics of the flowers. The data set consists of 150 cases, 50 for each class. The four attributes are Sepal Length, Sepal Width, Petal Length, and Petal Width.

The attributed that already been predicted belongs to the class of IRIS plant. The list of attributes present in the IRIS can be described as categorical, nominal, and continuous. The experts have mentioned that there isn't any missing value found in any attribute of this data set. The data set is complete. This project makes use of the well-known IRIS dataset, which refers to three classes of 50 instances each, where each class refers to a type of IRIS plant. The first of the classes is linearly distinguishable from the remaining two, with the second two not being linearly separable from each other. The 150 instances, which are equally separated between the three classes, contain the following four numeric attributes:

1. sepal length – continuous

2. sepal width – continuous

3. petal length – continuous

4. petal width – continuous

and the fifth attribute is the predictive attributes which is the class attribute that means each instance also includes an identifying class name, each of which is one of the following: IRIS Setosa, IRIS Versicolour, or IRIS Virginica the IRIS dataset (downloaded from the UCI repository, www.ics.uci.edu, which is a 150×4 matrix, is taken as the input data) (UCI, 1988).

In the analysis part, we first convert all the attributes to binary attributes as Table 6 below.

Table 7. Binary Attributes of IRIS Dataset

						Attribute-1			Attribute-2			Attribute-3			Attribute-4			
	Att-1	Att-2	Att-3	Att-4	Class	V-1	V-2	V-3	V-1	V-2	V-3	V-1	V-2	V-3	V-1	V-2	V-3	
1	5.1	3.5	1.4	0.2	Iris-setosa	1	0	0	0	0	1	1	0	0	1	0	0	Setosa
2	4.9	3	1.4	0.2	Iris-setosa	1	0	0	0	1	0	1	0	0	1	0	0	Setosa
3	4.7	3.2	1.3	0.2	Iris-setosa	1	0	0	0	0	1	1	0	0	1	0	0	Setosa
4	4.6	3.1	1.5	0.2	Iris-setosa	1	0	0	0	0	1	1	0	0	1	0	0	Setosa
5	5	3.6	1.4	0.2	Iris-setosa	1	0	0	0	0	1	1	0	0	1	0	0	Setosa
6	5.4	3.9	1.7	0.4	Iris-setosa	0	1	0	0	0	1	0	1	0	0	1	0	Versicolor
54	5.5	2.3	4	1.3	Iris-versicolor	0	1	0	1	0	0	0	1	0	0	1	0	Versicolor
55	6.5	2.8	4.6	1.5	Iris-versicolor	0	0	1	1	0	0	0	0	1	0	0	1	virginia
56	5.7	2.8	4.5	1.3	Iris-versicolor	0	1	0	1	0	0	0	0	1	0	1	0	Versicolor
104	6.3	2.9	5.6	1.8	Iris-virginica	0	0	1	0	1	0	0	0	1	0	0	1	virginia
105	6.5	3	5.8	2.2	Iris-virginica	0	0	1	0	1	0	0	0	1	0	0	1	virginia
106	7.6	3	6.6	2.1	Iris-virginica	0	0	1	0	1	0	0	0	1	0	0	1	virginia

As binarization is the process of transforming data features of any entity into vectors of binary numbers to make classifier algorithms more efficient, we now transform all the attributes of the dataset into binary vectors to represent all the data in the dataset as binary attributes. After converting all the data to suitable binary vectors based on each attribute, we run the proposed model in order to measure the performance of it and compare our model with existing classification models, SVM, k-NN and decision trees using IRIS dataset. Matlab_R2016b version program was used to compare these studies. The performance results are shown on Table 8.

Table 8. Comparison of Classification Models with Proposed Approach

Classification Method	Accuracy Rate (%)	# of Correctly Classified	# of Uncorrectly Classified	Processing Time (sec)
k-Nearest Neighbors	82%	123	27	4.18
Decision Tree	79%	118	32	3.48
Support Vector Machine	86%	129	21	4.36
Binarization Method	**86%**	**129**	**21**	**3.21**

As can be seen in Table-8, the proposed model we defined performed well on IRIS dataset. In general, our model gives a high accuracy rate of 86% where k-NN is around 82%, and decision trees gives an accuracy rate of 79%. The results show that our model gave the same results as SVM and gave better performance than k-NN and decision tree models.

5. Discussion and Conclusion

In real life, most data is processed as numeric or categorical data. The fact that the data is in this form requires that they be used together in the analyzes. Quantitative and qualitative data become interdependent in addressing common research questions and hypotheses. Meaningful integration allows researchers to realize the true benefits of mixed methods to produce a whole through integration that is greater than the sum of the individual qualitative and quantitative parts.

This paper presented a novel binary classification approach specifically designed for multi-class mixed data classification. The motivation behind this research stemmed from the prevalent nature of numeric and categorical data in real-life scenarios, necessitating their joint utilization in analyses. By integrating quantitative and qualitative data, researchers can effectively address common research questions and hypotheses, thereby harnessing the true potential of mixed methods.

The proposed approach in this study introduced a binarization algorithm to identify the classes of unclassified instances, along with the conversion of the dataset into a binary value matrix to handle large data sets efficiently. The model's performance was evaluated using a testing dataset, showcasing high classification accuracies across various cases and data types, including integer, discrete, continuous, binary, and categorical. Notably, the computational time and cost associated with the model were deemed acceptable, making it suitable for tackling large mixed data classification problems.

Additionally, the testing algorithm demonstrated computational tractability even for high-dimensional datasets, with reasonable total computational time observed in the examined scenarios. The accuracy values obtained by the proposed approach surpassed or were on par with other models, such as NN, SVM, Decision Trees, K-Nearest Neighbor, Logistic Regression, and Bayesian Classifier.

In summary, the development of this new approach has provided a viable solution to multi-class mixed data classification problems, leading to improved prediction accuracies. The proposed model not only exhibits high accuracy but also offers simplicity and understandability, making it a favorable choice for researchers and practitioners alike. By embracing this novel approach, the field of mixed data classification can continue to advance and yield valuable insights from diverse types of data.

Acknowledgement

The authors would like to thank the anonymous reviewers and editor for their comments and suggestions.

Conflict of Interest

No conflict of interest was declared by the authors.

References

Bai, Jing, Anran Yuan, Zhu Xiao, Huaji Zhou, Dingchen Wang, Hongbo Jiang, and Licheng Jiao. 2020. "Class Incremental Learning With Few-Shots Based on Linear Programming for Hyperspectral Image Classification." IEEE Transactions on Cybernetics.

Buniyamin, Norlida, Usamah bin Mat, and Pauziah Mohd Arshad. 2016. "Educational Data Mining for Prediction and Classification of Engineering Students Achievement." In 2015 IEEE 7th International Conference on Engineering Education, ICEED 2015, 49–53. Institute of Electrical and Electronics Engineers Inc.

Carnevalli, Jose A., and Paulo Cauchick Miguel. 2008. "Review, Analysis and Classification of the Literature on QFD-Types of Research, Difficulties and Benefits." International Journal of Production Economics 114 (2): 737–54.

Cover, T. M., and P. E. Hart. 1967. "Nearest Neighbor Pattern Classification." IEEE Transactions on Information Theory 13 (1): 21–27.

Dainotti, A., Pescape, A., claffy, k. (2012). Issues and Future Directions In Traffic Classification. IEEE Network, 1(26), 35-40.

E. Fix, J. J. Hodges, (1951) Discriminatory analysis: non-parametric discrimination: Consistency properties. Report No. 4, USAF School of Aviation Medicine, Randolph Field, TX, (1951).

Faes, L, M K Schmid, S K Wagner Bmbch, Liu Mbchb, R Chopra Bsc, N Pontikos, Sim Phd, et al. 2019. "Automated Deep Learning Design for Medical Image Classification by Health-Care Professionals with No Coding Experience: A Feasibility Study." Articles Lancet Digital Health 1: 232–74.

Fisher, R. A. 1986. "THE USE OF MULTIPLE MEASUREMENTS IN TAXONOMIC PROBLEMS." The Annals of Human Genetics, September, 179–88.

Fisher, R.A. 1988. "Iris Data Set." UCI Center for Machine Learning and Intelligent Systems. July 1, 1988. https://archive.ics.uci.edu/ml/datasets/Iris.

Ghamisi, P., Plaza, J., Chen, Y., Li, J., Plaza, A. (2017). Advanced Spectral Classifiers For Hyperspectral Images: a Review. IEEE Geosci. Remote Sens. Mag., 1(5), 8-32.

Graur, Dan, Yichen Zheng, and Ricardo B.R. Azevedo. 2015. "An Evolutionary Classification of Genomic Function." Genome Biology and Evolution 7 (3): 642–45.

Gupta, A. (2020). Approaches and Applications Of Early Classification Of Time Series: A Review.

Ho, W., Xu, X., Dey, P. (2010). Multi-criteria Decision Making Approaches For Supplier Evaluation and Selection: A Literature Review. European Journal of Operational Research, 1(202), 16-24.

Jouini, Mouna, Latifa Ben Arfa Rabai, and Anis ben Aissa. 2014. "Classification of Security Threats in Information Systems." In Procedia Computer Science, 32:489–96. Elsevier B.V.

Koggalahewa, D., Xu, Y., Foo, E. (2021). Unsupervised Spammer Detection In Social Networks Based On User Information Interests.

Kumar, V., Rahman, Z., Kazmi, A. (2016). Stakeholder Identification and Classification: A Sustainability Marketing Perspective. Management Research Review, 1(39), 35-61.

L'Hermitte, C., Tatham, P., Bowles, M. (2014). Classifying Logistics-relevant Disasters: Conceptual Model and Empirical Illustration. Journal of Humanitarian Logistics and Supply Chain Management, 2(4), 155-178.

Loh, Wei Yin. 2011. "Classification and Regression Trees." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 1 (1): 14–23.

Melin, Patricia, Frumen Olivas, Oscar Castillo, Fevrier Valdez, Jose Soria, and Mario Valdez. 2013. "Optimal Design of Fuzzy Classification Systems Using PSO with Dynamic Parameter Adaptation through Fuzzy Logic." Expert Systems with Applications 40 (8): 3196–3206.

Pal, Mahesh, and Giles M. Foody. 2010. "Feature Selection for Classification of Hyperspectral Data by SVM." IEEE Transactions on Geoscience and Remote Sensing 48 (5): 2297–2307.

Pratikakis, I, F Dupont, and M Ovsjanikov. 2017. "Exploiting the PANORAMA Representation for Convolutional Neural Network Classification and Retrieval." Eurographics Workshop on 3D Object Retrieval.

Russo, Mohammad Ashraf, Laksono Kurnianggoro, and Kang-Hyun Jo. 2019. 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE). IEEE.

Schwenker, Friedhelm, and Edmondo Trentin. 2014. "Pattern Classification and Clustering: A Review of Partially Supervised Learning Approaches." Pattern Recognition Letters 37 (1): 4–14.

Silva, Thiago Christiano, and Liang Zhao. 2012. "Network-Based High Level Data Classification." IEEE Transactions on Neural Networks and Learning Systems 23 (6): 954–70.

Singhal, Piyush, Gopal Agarwal, and Murali Lal Mittal. 2011. "Supply Chain Risk Management: Review, Classification and Future Research Directions." Journal of Business Science and Applied Management. Vol. 6.

Sutcu, Muhammed. 2020. "Effects of Total Cost of Ownership on Automobile Purchasing Decisions." Transportation Letters 12 (1): 18–24.

T. M. Cover, and P. E. Hart, "Nearest Neighbor Pattern Classification", Institute of Electrical and Electronics Engineers Transactions on Information Theory, 13 (1) (1967), 21-27.

UC Irvine. 2007. "UCI Machine Learning Repository." UCI. 2007. https://archive.ics.uci.edu/ml/datasets.php.

University of Toronto. 2003. "Data for Evaluating Learning in Valid Experiments." 2003. https://www.cs.toronto.edu/~delve/.

Varese, E., Lombardi, M. (2020). Dry Port: a Review On Concept, Classification, Functionalities And Technological Processes. Logistics, 4(4), 29.

Waltman, Ludo, and Nees Jan van Eck. 2012. "A New Methodology for Constructing a Publication-Level Classification System of Science." Journal of the American Society for Information Science and Technology 63 (12): 2378–92.

Webber, C., Gospodarowicz, M., Sobin, L., Wittekind, C., Greene, F., Mason, M., … & Groome, P. (2014). Improving the Tnm Classification: Findings From A 10-year Continuous Literature Review. Int. J. Cancer, 2(135), 371-378.

Wolfers, T., Floris, D., Dinga, R., Rooij, D., Isakoglou, C., Kia, S., … & Beckmann, C. (2019). From Pattern Classification To Stratification: Towards Conceptualizing the Heterogeneity Of Autism Spectrum Disorder. Neuroscience & Biobehavioral Reviews, (104), 240-254.

Zhang, Liang, Lingling Zhang, Weili Teng, and Yibing Chen. 2013. "Based on Information Fusion Technique with Data Mining in the Application of Finance Early-Warning." In Procedia Computer Science, 17:695–703. Elsevier B.V.

Zhang, Shichao, Xuelong Li, Ming Zong, Xiaofeng Zhu, and Ruili Wang. 2018. "Efficient KNN Classification with Different Numbers of Nearest Neighbors." IEEE Transactions on Neural Networks and Learning Systems 29 (5): 1774–85.