International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Systematic Literature Review of Mixed Variables Classification

Penny Ngu Ai Huong*[1], Hashibah Hamid[2]

Submitted:26/03/2024 **Revised**: 07/05/2024 **Accepted**: 18/05/2024

Abstract: Classification has become a widely used methodology across various fields. Numerous techniques are available for categorizing objects into distinct classes. Among these, the parametric approach, specifically the location model and its smoothed counterpart, has gained significant attention, particularly for data containing a mix of continuous and categorical variables. This paper systematically reviews literature from Scopus and Science Direct databases, focusing on a parametric classification approach involving mixed variables, i.e., location model and smoothed location model. A total of 70 articles were selected through systematic review procedures aligned with relevant thematic areas. These articles were analysed and discussed in the context of mixed classification based on the location model and smoothed location model across diverse data scenarios. Systematic literature reviews offer several advantages over traditional methods, notably providing a structured review process and essential priorities to mitigate research biases effectively. The review reveals that the smoothed location model demonstrates superior classification performance compared to other techniques in most studies. However, literature specifically addressing the applications of the location model and smoothed location model for classifying objects with mixed variables remains limited. Consequently, it is recommended that future research endeavours consider both models when dealing with classification tasks involving mixtures of variables.

Keywords: *Classification; Location Model; Smoothed Location Model; Mixed Variables; Systematic Literature Review.*

1. Introduction

The categorisation of data comprising both continuous and discrete variables has gained significant attention in recent decades. The advancement of statistical inference techniques for analysing such mixed data holds considerable importance, given its widespread applications across various research domains, including medical, physiological, and social sciences. Furthermore, experimental data are often collected to forecast the values of one or more system attributes (responses), which may manifest in quantitative or qualitative formats. For instance, in medical diagnostics and prognostics, extensive investigations involve data collection incorporating a blend of discrete and continuous factors (such as health status, disease severity, and prognosis outcomes) to allocate patients into appropriate diagnostic or prognostic categories.

Currently, various techniques exist for analyzing random variables that include both continuous and discrete components. For instance, classification methodologies are employed to assess genetic diversity among accessions within gene banks, wherein accessions are categorised into sub-populations or clusters based on attributes such as plant morphology, agronomic performance, disease resistance, genetic markers, and other pertinent factors [1]. Within the realm of omics, numerous interesting challenges arise where the response to be predicted assumes a qualitative nature, characterized by a discrete set of values [2]. For instance, discerning the healthiness of individuals based on experimental data exemplifies such a scenario. Here, the qualitative response to be anticipated is represented by the two discrete values, "healthy" and "unhealthy".

Using the gathered experimental data, classification methodologies serve as appropriate tools for constructing models to predict the most suitable class to which individuals under investigation belong [2]. Software defect prediction entails a quality control process that leverages past defect data alongside software parameters. Before the software testing phase, this predictive technique facilitates the identification of defect-prone software modules. The predominant approach in research involves applications of machine learning classification techniques to categorise software modules into either defect-prone or non-defect-prone classifications [3]. A study by Kaya et al. [4] employed various classification methods including AdaBoostM1, linear discriminant, support vector machine, random forest, subspace discriminant, and weighted *k*nn as these methods are widely used in defect predicting.

Typically employed for separating classes with predefined categories, discriminant analysis is a widely utilized data analysis method, in which the dependent variables are categorical, and the independent variables are quantitative and exhibit a normal distribution [5,6]. One prominent model for analyzing mixed data is the general location model (GLOM), often utilized in discriminant analysis when both qualitative (discrete) and quantitative (continuous) data are involved in the independent variables. Initially proposed by Olkin and Tate [7], the location model

[1] *Asia Pacific University of Technology & Innovation (APU) – 57000 Malaysia*
ORCID ID: 0009-0001-2805-1999
[2] *Universiti Utara Malaysia – 06010, MALAYSIA*
ORCID ID: 0000-0002-7908-9495
** Corresponding Author Email: pennyngu90@hotmail.com, penny.ngu@apu.edu.my*

assumes a conditional distribution of continuous variables, assuming multivariate normally distributed discrete variables with a constant covariance matrix across all cells indicated by the discrete variables [8]. Chang and Afifi [9] extended the application of the location model to two-classes scenarios, developing a Bayes classification method for categorising observations containing both continuous and dichotomous variables. Krzanowski [10] expanded on these findings, deriving optimal and estimated allocation rules for mixed binary and continuous variables using likelihood ratio. Substantial progress has been achieved in various aspects including expected error rate computation, variable selection, heteroscedasticity of between-population dispersion, and heteroscedasticity of cross-location dispersion [11-14]. Krzanowski [12] offers an overview of the advancements from the location model and outlines prospects. In another study, populations are classified using mixed covariates comprising discrete and continuous variables, assuming homogeneous conditional dispersion matrices between two populations specific to the discrete values [15].

In discriminatory contexts using the location model, it is crucial to restrict the number of discrete variables to avoid an excessive number of parameters to be estimated. When initial sample sizes from each class are small, Krzanowski [16] recommended employing a maximum of six binary variables, reducing this number further when variables possess more than two states. The computational effort required for estimating misclassification rates increases significantly with the number of binary variables, posing challenges particularly when initial sample sizes are substantial. Hence, Krzanowski [16] advocated a discrete variable selection approach involving backward elimination to identify a suitable reduced location model for discriminant applications featuring many binary variables.

Conversely, the traditional maximum likelihood estimation in the location model faces challenges when empty cells arise due to numerous binary variables measured in the study. To address this issue, a non-parametric smoothing approach is proposed to estimate parameters within the location model [17]. Additionally, concerns regarding over-parameterization and instability of the covariance matrix in the location model have been tackled by some researchers through the integration of non-parametric smoothing and regularization techniques [18]. More recently, Hamid [19] introduced a concept that combines principal component analysis for dimensionality reduction with a discriminant function based on the location model. The objective of her study is to offer an alternative classification strategy for situations where the observed variables are mixed and exceptionally large.

Numerous contemporary studies worldwide are focused on classification methods. As earlier mentioned, using mixed variables in classification is common across various disciplines. Consequently, this paper presents a systematic review of classification methods to elucidate their benefits across diverse fields further. Moreover, there is a notable shortage of studies investigating the applications of the location model and smoothed location model in mixed variables classification. Therefore, the subsequent section reviews the methodology concerning mixed variables classification purposes based on both location models.

2. Materials and Methods

This section provides comprehensive details, with all procedures thoroughly explained. It is organized into several subsections as outlined below.

2.1. Identification

The systematic review process involves three primary phases to select relevant papers for this study. Initially, keywords are identified, and related terms are sought using various resources such as thesauruses, dictionaries, encyclopaedias, and prior research. Once all relevant keywords are identified, search strings are formulated and applied in the Scopus and Science Direct databases (see Table 1). In the initial stage of the systematic review, 152 papers were retrieved from both databases for the current research project.

Table 1. The search strings

Scopus	TITLE-ABS-KEY (“classification” AND “classifications” AND "locations model" AND "location models") AND (LIMIT-TO (DOCTYPE, "ar")) AND (LIMIT-TO (LANGUAGE, "English"))
Science Direct	Title, abstract, keywords: “classification” AND “classifications” AND "locations model" AND "location models"

2.2. Screening

In the initial screening stage, duplicate articles are eliminated. Subsequently, researchers gained 136 articles based on the predefined inclusion and exclusion criteria, which resulted in the exclusion of 16 articles. Primary literature, such as research articles, served as the foremost criterion due to its significance as a source of valuable knowledge. Additionally, publications such as systematic reviews, reviews, meta-analyses, and meta-syntheses were also considered. However, publications like books, book series, chapters, conference proceedings, and articles in non-English languages were excluded from this study. These specific screening criteria resulted in the exclusion of 45 articles.

2.3. Eligibility

The third stage, termed eligibility, currently comprises only 91 articles, which are deemed suitable. Each article title and its substantial content underwent meticulous scrutiny to ensure alignment with the inclusion criteria and objectives of the present study. Subsequently, 21 articles were excluded as they did not pertain to the designated topics. As a result, 70 articles remained available for review (see Table 2).

Table 2. The selection criterion is searching

Criterion	Inclusion	Exclusion
Language	English	Non-English
Literature type	Journal (research articles only)	Journal (books, book series, book chapters, conference proceedings)

2.4. Data Abstraction and Analysis

In this study, an integrative analysis was conducted to analyse and synthesize several research designs including qualitative, quantitative, and mixed methodologies as depicted in Figure 1.

3. Results

In many real-world contexts, classification is prevalent, spanning fields such as medicine, manufacturing, and beyond. Through our search methodologies, we identified and analyzed 70 articles. These articles were categorized into three primary themes: classification (36 articles), location model (32 articles), and smoothed location model (2 articles).

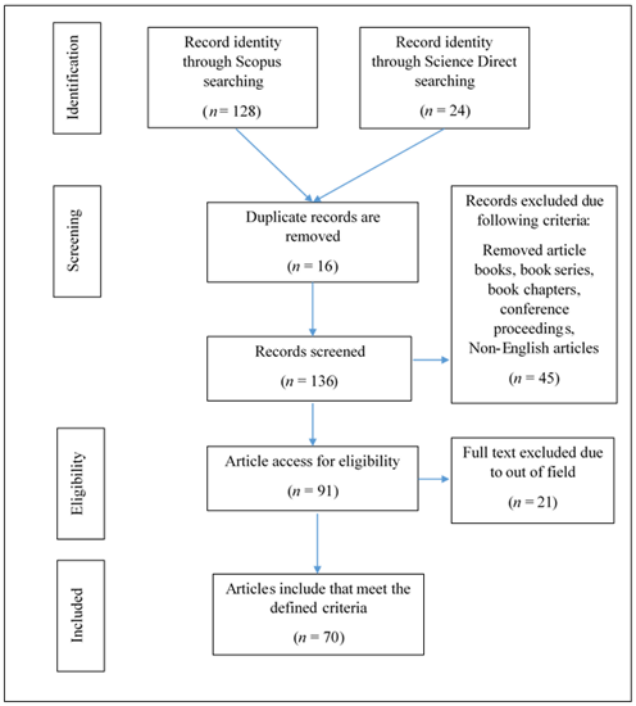

Fig. 1. Flow diagram of the proposed search study

3. Results

In many real-world contexts, classification is prevalent, spanning fields such as medicine, manufacturing, and beyond. Through our search methodologies, we identified and analysed 70 articles. These articles were categorised into three primary themes: classification (36 articles), location model (32 articles), and smoothed location model (2 articles).

3.1. Classification

One of the most significant ongoing discussions centres on classification, which entails the assignment of each entity to a distinct class within a set of mutually exclusive and non-overlapping classes. Within the theory of library and information science (LIS), classification encompasses three interconnected yet different concepts. Firstly, it denotes a system of classes utilized to categorise and structure diverse activities based on predetermined guiding principles. Secondly, it refers to individual classes within a categorisation system. Lastly, classification encompasses the action of categorising objects into classes within a classification system [20].

A considerable body of published research highlights the utility of classification across various domains. In agriculture, classification has been applied in categorising maize races and common beans. For instance, one study employed a numerical classification technique to categorise eight Peruvian highland maize races based on six vegetative features assessed over two years, comparing the employed classification to the existing racial classification [21]. Another study compares different methods for the classification of gene bean accessions [22]. This

classification approach suggests that the collection of bean accessions could serve as a valuable resource for plant breeding due to its basis in genetic diversity. In a separate investigation, the genetic parameters and the restricted maximum likelihood (REML) method, or the best linear unbiased prediction (BLUP), were assessed for their effectiveness in predicting the genetic values of Vitis [23]. Three homogeneous sets were formed to identify Vitis genotype classes based on hybrid resistance and broad-sense heritability values. Additionally, regression analysis was developed as a classification method to predict the durability of structural wood members in fire [24]. The durability of these members depends on the cross-sectional characteristics induced by wood charring. Factors such as char contraction, density, and moisture content influence the wood's charring rate, emphasizing the significance of moisture content and surface recession of wood charring.

In addition to agriculture, numerous studies have investigated the application of classification in categorising animal species. For instance, species distribution modelling (SDM) methods such as classification and regression trees (CART) and boosted regression trees (BRT) have been utilized to re-evaluate the environmental factors influencing the global distribution of coral reefs [25]. In another study, logistic regression (LR) and classification tree (CT) models were employed to estimate microhabitat utilization and the summer distribution of juvenile Atlantic salmon [26]. These models predicted the presence or absence of salmon at specific locations based on the observed habitat variables such as depth, current velocity, substratum particle size, distance to stream bank, presence of instream and overhead covered. To predict the species rarity of reptiles and amphibians in southern California, data on species presence and absence from 420 locations were utilized to develop classification trees, generalized additive models, and generalized linear models. The performance of the models was assessed using sensitivity, specificity, and the area under the curve (AUC) of the receiver-operating characteristic (ROC) plot based on twofold cross-validation or bootstrapping. Predictors included climate, geography, soil, and vegetation characteristics. The findings indicated that employing a variety of modelling techniques resulted in more accurate species distribution models for rarer species [27].

Numerous scholars have observed a growing body of literature concerning the utilization of classification methods across various analytical contexts. For instance, support vector machines (SVM), k-nearest neighbours (*k*-nn), naive Bayes algorithms, regression, naïve Bayes, cluster analysis, and convolutional neural networks. In applications with Twitter data, researchers have employed classification techniques such as SVM, *k*-nn, and naive Bayes algorithms to detect earthquakes swiftly. Subsequently, tweets are analysed to identify individuals linked to the event, and requests for aid and volunteer offers are collated [28]. In another study by Strath et al. [29], time series data were discretized into bins and analysed using SVM for activity classification and Bagged decision tree regression for metabolic cost prediction. Furthermore, two supervised classificatory methods, maximum likelihood classification (MLC) and SVM classification were employed for image classification [30].

In 2021, Colaço and Abreu [31] introduced a novel categorisation framework relying on cluster analysis (specifically *k*-means) and a limited set of variables including density, diversity, and clustering. This framework was implemented by Lisbon in three different years (1995, 2002 and 2010), which can be construed as a classification system. It was established through a cross-sectional examination of commercial structures, illustrating its effectiveness in depicting commercial locations and changes over time. The correlation between commercial classification and location modelling may be strengthened, featuring the significance of commercial studies in urban planning and policymaking. Notably, the minimal dataset employed also allows for the utilization of cluster membership in location models. This study provides valuable insights, indicating that cluster analysis remains a focal point of interest for researchers. Additionally, another study utilized classification techniques for social media text streams during emergencies, demonstrating the model's capability to categorise such streams based on the topics they addressed [32].

Distribution lines constitute a vital component of modern power systems, directly influencing the stability and security of power supply to prevent further financial and social costs stemming from load interruptions. A robust power system protection program should swiftly identify all instances of defects. Fault diagnosis entails two primary tasks: fault classification, which has achieved notable accuracy rates, and fault location. Two studies were conducted addressing fault diagnosis using neural networks (NN) and convolutional NN. A study inspired by the Fourier transforms proposed an online data-driven method that converts signals from the time domain to the image domain through the signal-to-image (SIG) algorithm. Subsequently, these converted images are processed using a convolutional NN architecture, notable for its compactness compared to others, declaring it easier to transplant onto hardware platforms and requiring less memory space [33]. In another study, Ferreira et al. [34] suggested the utilization of autonomous NN to establish a correlation between electrical signals at one terminal and transmission line defect information. This method yields an error range around the predicted short-circuit position.

In clinical diagnostics, mass spectrometric techniques such as gas chromatography-mass spectrometry, liquid

chromatography-mass spectrometry, and matrix-assisted laser desorption ionization - time-of-flight mass spectrometry (MALDI-TOF MS) are commonly employed to detect large biomolecules present in trace amounts [35]. In an article published in the Journal of Microbiological Methods, a panel assessing the value of MALDI-TOF MS biomarker peaks and their correlation to outbreak strains, locations, sources, patients, diagnoses, and isolate genetics was constructed using 55 well-characterized Burkholderia contaminants isolates [36]. Unsupervised clustering techniques were employed, and classification models were generated utilizing biostatistical analysis software. The findings indicated that MALDI-TOF MS can identify isolates that may not be typable by the pulsed-field gel electrophoresis (PFGE) and effectively differentiate Burkholderia contaminants isolates into clonal, epidemiological clusters. Further research is necessary to enhance understanding of this capability.

3.2. Location Model

The linear discriminant function (LDF), originally introduced by Fisher [37], is commonly employed for discriminating and classifying new objects into one of two populations. However, some users confined the application of this method solely to either discrete or continuous variables. In 1975, Krzanowski [10] proposed a classification rule based on a location model (LM) capable of handling binary and continuous variables. This proposed rule underwent evaluation and application across various datasets, demonstrating satisfactory classification performance compared to the LDF. Furthermore, the author expanded the research by examining the LM through the computation of Mahalanobis distances between classes and visualization of resulting configurations using principal coordinate analysis on breast cancer data in 1976. In 1977, Krzanowski [38] assessed the performance of the LDF when certain assumptions of multivariate normality were violated, including unequal variance-covariance matrices and non-normality of data. The study suggested adapting the location model for mixed binary and continuous data in cases where the LDF exhibits poor classification performance. Additionally, in 1980, Moussa [39] proposed a linear additive model to estimate parameters for the LM when the data comprised both binary and continuous variables with zero frequency or objects in certain states.

Prior research has documented discrimination between two populations when data comprises mixtures of binary and continuous variables. Consequently, in 1986, Krzanowski [40] expanded the concept of discriminant rule from two populations to encompass more than two populations. This proposed rule was subsequently compared with the traditional normal-based rule. In 1993, Krzanowski [12] authored a paper primarily to investigate the extensive advancements and capabilities of the LM, focusing on inter-class distances, discriminant analysis, error rates, estimation, feature selection, and management of missing data. In the subsequent year, 1994, Krzanowski [13] further explored the assumptions underlying the LM by relaxing the common within-cell dispersion matrices to allow discrimination of different matrices between two populations. This alteration transitioned the Bayes location from a choice among linear functions to a choice of quadratic functions.

Several researchers have expanded the framework of the LM. They developed a predictive allocation rule for classification, which relies on the standard frequency distribution of the LM and a vague prior distribution for the unknown parameters [41]. The author compared the performance of the predictive rule and its estimative rule in two scenarios: (1) the binary variables failed to discriminate between the two populations, and (2) the discrimination exists between the populations. The findings indicated that the predictive rule yielded a lower misclassification rate in the former case, while the estimative rule performed well in the latter. In another study, Willse and Boik [42] demonstrated that Lawrence and Krzanowski's [43] finite mixture model cannot be identified without imposing further constraints. Consequently, the authors proposed identifiable finite mixture models by restricting the conditional means of the continuous variables. Simulations were employed to evaluate these newly proposed models. The conditional mean structure of the continuous variables in the restricted location mixture models resembles Everitt's [44] underlying variable mixture models, but the restricted location mixture models offer computational advantages.

Later, Leung [15] examined the classification of mixed discrete and continuous variables within the framework of the LM, considering mixed covariates. Some variables termed covariates, lack discriminative power in classification due to their equivalent means across classes. These covariates can be leveraged to generate a new variable to improve classification [45]. A plug-in version of the Bayes rule is utilized, incorporating complete covariate adjustment for classification. Regularization within the general LM is suggested when the ratio of dimensionality of continuous variables to the total training sample approaches unity but remains slightly less [46]. Under specific conditions, a limiting overall expected error for the classifier is provided, and this error can guide the determination of optimized regularization parameters [14].

The general location model (GLOM) and the conditional grouped continuous model (CGCM) have been commonly employed in practice to handle mixed variables in discrimination analysis, treating ordinal variables as nominal and nominal variables as ordinal, respectively [47]. However, these approaches may lead to information loss when ordinal variables are categorised as nominal and raise

concerns about model robustness when treating ordinal variables as continuous [48]. Consequently, a new approach, the general mixed-data models (GMDMs), has been proposed to handle mixed data including nominal, ordinal, and continuous variables. Research demonstrated that GMDMs offer satisfactory classification performance, as evidenced in a childhood croup study [49]. Traditionally, GLOM assumes that continuous multivariate distributions across cells are generated by various combinations of categorical variables with equal covariance matrices. In 2017, Amiri et al. [50] introduced an extension of GLOM that accounted for both equal and unequal covariance matrices. Three covariance structures; same factor analyser, factor analyser with unequal specific variance matrices (in general and parsimonious forms), and factor analyser with shared factor loadings, which are employed across cells for modelling covariance structure and parameter reduction. Illustrative of real-data analyses highlighted the effectiveness of the classification performance of these models.

Restricting the number of discrete variables becomes imperative when utilizing the LM for discriminatory analyses to prevent an excessive number of parameters requiring estimation. The computational burden associated with estimating misclassification rates was observed to grow exponentially with the number of discrete variables, posing challenges particularly when dealing with large initial sample sizes. Baah et al. [51] investigated scenarios where the number of binary variables is proportionate to the number of continuous variables, as parameter estimation in the LM relies on the multinomial cells formed by the discrete variables. The study aimed to identify an optimal ratio of the continuous variables to the binary, minimizing the misclassification error rates in the two-class scenario.

Various researchers have conducted several studies exploring location models. For instance, one study compares the effectiveness of the LM approach in discriminant analysis with a method established in rough sets theory using a real medical dataset [52]. Additionally, another investigation assesses the performance of five multiple imputations with chained equations (MICE) techniques for handling imperfect nominal variables [53]. These techniques include MICE utilizing polytomous regression for elementary imputation, CART, nested logistic regressions, Allison's [54] ranking procedure, and a joint modelling approach based on the general LM. In most conditions, Allison's [54] ranking method and MICE with CART showed poor performance, while MICE with polytomous regression exhibited superior performance.

The LM operates under the assumption that objects exist in all potential combinations of multinomial variable values and subpopulations. However, practical applications often encounter situations where certain multinomial cells remain unoccupied. To address this, Franco et al. [55] introduced the modified location model (MLM) alongside Ward's method, implementing a two-stage clustering strategy. Their findings suggest that MLM can effectively analyse datasets containing empty cells. Employing a two-stage approach involving the Ward method to identify primary class and MLM to refine class composition appears promising. The authors expanded their investigation from two-way to three-way data, incorporating categorical and continuous variables into MLM. This strategy was evaluated by classifying objects in simulated datasets with known structures and two experimental datasets consisting of multi-attributes for classifying genetic resources in multiple environments [56]. Moreover, the three-way Ward-MLM clustering strategy was applied across three distinct datasets to assess its efficacy in grouping cultivars with low imperfect genotypic correlation (COI) [57].

Many researchers have sought to elucidate the application of Ward-MLM methodology in classifications of gene bean accessions, genotypes, or landraces. For instance, Ward-MLM methodology has been utilized to categorise gene accessions, classify landraces into five clusters, identify redundant landraces, leading to a reduction in the number of accessions in subsequent critical trials, and class 24 accessions of related Peruvian highland maize races, among other applications [58-64].

The LM encounters challenges due to empty cells arising from the measured variables, notably the binary one. Excessive binary variables coupled to numerous empty cells may lead to inaccurate parameter estimates. The study advocates for a variable selection technique predicated on class distance, employing smoothed Kullback-Leibler divergence pooled with the LM. Later, a novel concept is proposed, integrating principal component analysis (PCA) for dimensionality reduction with a discriminant function based on the LM [19]. The study aims to offer practitioners an alternative effective tool for addressing classification dilemmas in cases where observed variables are diverse and abundant. Additionally, the author recommended incorporating nonlinear PCA into the classical location model (cLM) to handle numerous categorical variables, thereby mitigating the high misclassification rates [65]. This inquiry validates the efficacy of the proposed model's novel discrimination technique as a viable solution to classification challenges associated with mixed variables, especially when struggling with an excess of categorical variables.

3.3. Smoothed Location Model

The location model (LM) is a prominent approach employed in discriminant analysis for multivariate datasets containing a mix of categorical and continuous variables. However, it encounters challenges of over-parameterization, particularly when an analysis involves numerous

parameters. This can necessitate constraints on the number of binary variables or require a substantial number of training instances [12,66]. Addressing this issue, a study recommended non-parametric smoothing techniques to expand the scope of applicability and significantly reduce the number of parameters requiring estimation [17]. Comparative analysis with other methods demonstrated that these newly proposed non-parametric smoothing techniques exhibited favourable performance.

The smoothed location model (SLM) is an effective method that can manage both continuous and binary variables concurrently [67]. Nonetheless, the SLM becomes impractical when confronted with a large number of mixed variables. To address this issue, a variable extraction technique, PCA, was combined with the SLM. The study's findings are promising, recommending the proposed method for managing large numbers of mixed variables.

In addition to PCA, another method for reducing the abundance of binary and continuous variables involved the applications of multiple correspondence analysis (MCA) in conjunction with PCA [68]. MCA encompasses four types: Indicator MCA, Burt MCA, Adjusted MCA, and Joint Correspondence Analysis (JCA). This study aims to construct novel SLMs by integrating SLM with two variable extraction techniques, namely PCA and two types of MCA, with the primary objective of reducing the surplus of mixed variables, particularly binary variables. The efficacy of the newly developed models is assessed based on the misclassification rate, with evaluation conducted on SLM+PCA+Indicator MCA and SLM+PCA+Burt MCA models. Results indicated that the novel SLMs combined with the two variable extraction methods, offer a viable solution in addressing issues of mixed variables in classification tasks, particularly when dealing with huge binary variables.

Nevertheless, the effectiveness of the location model reduces when the dataset contains outliers [69]. To mitigate the impacts of outliers in constructing the discriminant rules based on the LM, robust techniques are employed. Addressing both outliers and empty cells challenge concurrently, a novel LM is introduced in the research. Winsorization is incorporated into the smoothing method to tackle outliers in the mixed variables. Results affirmed that the newly devised methodology of LM provides practitioners with additional tools for addressing classification difficulties containing outliers and overcoming the LM's limitation concerning empty cells [70]. Statistical analyses revealed the optimal performance of the newly developed LM even with the presence of outliers. Furthermore, empirical findings validated the effectiveness and applicability of the proposed classification approach.

4. Discussion

In this systematic review, we assessed 135 journal articles, ultimately selecting 70 articles focusing on classification methods, location models, and smoothed location models. Our review provided insights into various classification methods applied across different fields. For example, SVM is commonly employed for image classification, whereas neural networks are utilized for fault diagnosis [30,33].

Based on the reviewed literature, the LM emerges as a prominent approach for handling mixed variables in classification tasks. However, when empty cells pose challenges for maximum likelihood parameter estimation, the SLM offers a viable solution [17]. Previous research highlighted that empty cells often arise once numerous binary variables are measured in the study [12]. Various strategies have been proposed in the literature to address this issue. For instance, integrating variable extraction or variable reduction techniques with the LM or SLM has been explored in several studies to tackle the challenges posed by a large consideration of the binary variables [19,64,68].

5. Conclusions

This systematic review comprehensively examined the literature concerning various classification methods, particularly focusing on the LM and SLM for mixed variables classification. The findings suggested that the SLM demonstrates improved classification performance compared to the alternative and other methods in most conditions. However, the number of articles specifically addressing the use of these models in mixed variables classification is limited. The shortage of studies on the SLM hampers further exploration of its efficacy. Hence, future research should prioritize investigations into the SLM, specifically when dealing with mixed variables for classification purposes.

Acknowledgements

This research was supported by the Ministry of Higher Education (MoHE) of Malaysia through the Fundamental Research Grant Scheme (FRGS/1/2019/STG06/UUM/02/5) with S/O code 14374.

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] J. Franco, J. Crossa, J. Villaseñor, A. Castillo, S. Taba, & S. A. Eberhart, “A two-stage, three-way method for classifying genetic resources in multiple environments,” *Crop Science*, vol. 39, no.1, pp. 259–267, 1999. doi:10.2135/cropsci1999.0011183X003900010040x

[2] M. Cocchi, A. Biancolillo, & F. Marini, “Chemometric Methods for Classification and Feature Selection,

Comprehensive Analytical Chemistry," vol. 82, no.1, pp. 265–299, 2018. doi:10.1016/bs.coac.2018.08.006

[3] C. Catal, & B. Diri, "A systematic review of software fault prediction studies. Expert Systems with Applications," vol. 36, no.4, pp. 7346–7354, 2009. doi:10.1016/j.eswa.2008.10.027

[4] A. Kaya, A. S. Keceli, C. Catal, & B. Tekinerdogan, "Model analytics for defect prediction based on design-level metrics and sampling techniques," In Model Management and Analytics for Large Scale Systems, pp. 125–139, 2020. doi:10.1016/B978-0-12-816649-9.00015-6

[5] P. Baah, A. Adebanji, & R. G. Kakaï, "Optimal ratio of continuous to categorical variables for the two-group location model," International *Journal of Applied Mathematics and Statistics*, vol. 42, no.12, pp. 18–26, 2013.

[6] H. Hamid, F, Zainon, & T. P. Yong, "Performance analysis: An integration of principal component analysis and linear discriminant analysis for a very large number of measured variables," *Research Journal of Applied Sciences*, vol. 11 no.11, pp. 1422-1426, 2016.

[7] I. Olkin & R. F. Tate, "Multivariate Correlation Models with Discrete and Continuous Variables," *The Annals of Mathematical Statistics*, vol. 32, pp. 448–465, 1961.

[8] G. J. McLachlan, "Discriminant Analysis and Statistical Pattern Recognition," New York, NY: John Wiley & Sons, Inc,1992.

[9] P. C. Chang, & A. A. Afifi, "Classification based on Dichotomous and Continuous Variables," Journal of the American Statistical Association, vol. 69, no.346, pp. 336–339, 1974.

[10] W. J. Krzanowski, "Discrimination and classification using both binary and continuous variables," *Journal of the American Statistical Association*, vol. 70, no.352, pp. 782–790, 1975. doi:10.1080/01621459.1975.10480303

[11] N. Balakrishnan, S. Kocherlakota, & K. Kocherlakota, "On the errors of misclassification based on dichotomous and normal variables," *Annals of the Institute of Statistical Mathematics*, vol. 38, no.3, pp. 529–538, 1986. doi:10.1007/BF02482540

[12] W. J. Krzanowski, "The location model for mixtures of categorical and continuous variables," *Journal of Classification*, vol. 10, no.1, pp. 25–49, 1993. doi:10.1007/BF02638452

[13] W. J. Krzanowski, "Quadratic location discriminant functions for mixed categorical and continuous data," *Statistics & Probability Letters*, vol. 19, no.2, pp. 91–95, 1994. doi.:10.1016/0167-7152(94)90138-4

[14] C. Y. Leung, "Regularized classification for mixed continuous and categorical variables under across-location heteroscedasticity," *Journal of Multivariate Analysis*, vol. 93, no.2, pp. 358–374, 2005. doi:https://doi.org/10.1016/j.jmva.2004.03.001

[15] C. Y. Leung, "Error rates in classification consisting of discrete and continuous variables in the presence of covariates," *Statistical Papers*, vol. 42, no.2, pp. 265–272, 2001. doi:10.1007/s003620100055

[16] W. J. Krzanowski, "Stepwise Location Model Choice in Mixed Variables in Discriminant Analysis," *Applied Statistics*, vol. 32, no.3, pp. 260–266, 1983.

[17] O. Asparoukhov, & W. J. Krzanowski, "Non-parametric smoothing of the location model in mixed variable discrimination," *Statistics and Computing*, vol. 10, no.4, pp. 289–297, 2000. doi:10.1023/A:1008973308264

[18] R. Guti´ errez, A. Merbouha, R. Guti´errez-S´anchez, & A. Nafidi, "Non-parametric smoothing and regularization of the location model in mixed variable discrimination," Monograf´ıas Del Seminario Matem´ Atico Garc´ıa de Galdeano, pp. 107–116, 2008.

[19] H. Hamid, "A new approach for classifying large number of mixed variables," *World Academy of Science, Engineering and Technology*, vol. 46, pp. 156–161, 2010.

[20] E. K. Jacob, "Classification and Categorization: A Difference that Makes a Difference," *Library Trends*, vol. 52, no. 3, pp. 515-540, 2004.

[21] R. Ortiz, R. Sevilla, G. Alvarado, & J. Crossa, "Numerical classification of related Peruvian highland maize races using internal ear traits," *Genetic Resources and Crop Evolution*, vol. 55, no.7, pp. 1055–1064, 2008. doi:10.1007/s10722-008-9312-3

[22] Z. Knezović, J. Gunjača, Z. Šatović, & I. Kolak, "Comparison of different methods for classification of gene bank accessions," *Agriculturae Conspectus Scientificus*, vol. 70, no.3, pp. 87–91, 2005.

[23] P. R. dos Santos, A. P. Viana, V. M. Gomes, S. da Costa Preisigke, O. F. de Almeida, E. A. Santos & M. A. Walker, "Resistance to Pratylenchus brachyurus in Vitis species population through multivariate approaches and mixed models," *Scientia Agricola*, vol. 76, no.5, pp. 424–433, 2019. doi:10.1590/1678-992x-2017-0387

[24] R. H. White & E. V. Nordheim, "Charring rate of wood for ASTM E 119 exposure," *Fire Technology*, vol. 28, no.1, pp. 5–30, 1992. doi:10.1007/BF01858049

[25] E. Couce, A. Ridgwell, & E. J. Hendy, "Environmental controls on the global distribution of shallow-water coral reefs," *Journal of Biogeography*, vol. 39, no.8, pp. 1508–1523, 2012. doi:10.1111/j.1365-2699.2012.02706.x

[26] K. Turgeon, & M. A. Rodríguez, "Predicting microhabitat selection in juvenile Atlantic salmon Salmo salar by the use of logistic regression and classification trees," *Freshwater Biology*, vol. 50, no.4, pp. 539–551, 2005. doi:10.1111/j.1365-2427.2005.01340.x

[27] J. Franklin, K. E. Wejnert, S. A. Hathaway, C. J. Rochester, & R. N. Fisher, "Effect of species rarity on the accuracy of species distribution models for reptiles and amphibians in southern California," *Diversity and Distributions*, vol. 15, no.1, pp. 167–177, 2009. doi:10.1111/j.1472-4642.2008.00536.x

[28] O. B. Gulesan, E. Anil, & P. S. Boluk, "Social media-based emergency management to detect earthquakes and organize civilian volunteers," *International Journal of Disaster Risk Reduction*, vol. 65, 10254, 2021. doi:https://doi.org/10.1016/j.ijdrr.2021.102543

[29] S. J. Strath, R. J. Kate, K. G. Keenan, W. A. Welch, & A. M. Swartz, "Ngram time series model to predict activity type and energy cost from wrist, hip and ankle accelerometers: Implications of age," *Physiological Measurement*, vol. 36, no.11, pp. 2335–2351, 2015. doi:10.1088/0967-3334/36/11/2335

[30] R. Banerjee & P. K. Srivastava, "Reconstruction of contested landscape: Detecting land cover transformation hosting cultural heritage sites from Central India using remote sensing," *Land Use Policy*, vol. 34, pp. 193–203, 2013. doi:https://doi.org/10.1016/j.landusepol.2013.03.005

[31] R. Colaço & J. de Abreu e Silva, "Commercial classification and location modelling: Integrating different perspectives on commercial location and structure," *Land*, vol. 10, no.6, 2021. doi:10.3390/land10060567

[32] Y. Wang, T. Wang, X. Ye, J. Zhu, & J. Lee, "Using social media for emergency response and urban sustainability: A case study of the 2012 Beijing rainstorm," *Sustainability (Switzerland)*, vol. 8, no.1, pp. 1–17, 2016. doi:10.3390/su8010025

[33] Y. Yu, M. Li, T. Ji, & Q. H. Wu, "Fault location in distribution system using convolutional neural network based on domain transformation," *CSEE Journal of Power and Energy Systems*, vol. 7, no.3, pp. 472–484, 2021. doi:10.17775/CSEEJPES.2020.01620

[34] V. H. Ferreira, R. Zanghi, M. Z. Fortes, S. Gomes, & A. P. Alves da Silva, "Probabilistic transmission line fault diagnosis using autonomous neural models," Electric Power Systems Research, vol. 185, 106360, 2020. doi:https://doi.org/10.1016/j.epsr.2020.106360

[35] Y.T. Cho, H. Su, W.J. Wu, D.C. Wu, M.F. Hou, C.H. Kuo, & J. Shiea, "Biomarker Characterization," by *MALDI–TOF/MS*, pp. 209–254, 2015. doi:10.1016/bs.acc.2015.01.001

[36] S. Fiamanya, L. Cipolla, M. Prieto, & J. Stelling, "Exploring the value of MALDI-TOF MS for the detection of clonal outbreaks of Burkholderia contaminans," *Journal of Microbiological Methods*, vol. 181, 106130, 2021. doi:https://doi.org/10.1016/j.mimet.2020.106130

[37] R. A. Fisher, "The Use of Multiple Measurements in Taxonomic Problems," *Annals of Eugenics*, vol. 7, no.1, pp. 179–188, 1936. doi:10.1007/s13398-014-0173-7.2

[38] W. J. Krzanowski, "The performance of fisher's linear discriminant function under non-optimal conditions," *Technometrics*, vol. 19, no.2, pp. 191–200, 1977. doi:10.1080/00401706.1977.10489527

[39] M. A. A. Moussa, "Discrimination and allocation using a mixture of discrete and continuous variables with some empty states," *Computer Programs in Biomedicine*, vol. 12, no. 2, pp. 161–171, 1980. doi:https://doi.org/10.1016/0010-468X(80)90062-8

[40] W. J. Krzanowski, "Multiple discriminant analysis in the presence of mixed continuous and categorical data," *Computers & Mathematics with Applications*, vol. 12, no.2, pp. 179–185, 1986 doi:https://doi.org/10.1016/0898-1221(86)90071-4

[41] I. G. Vlachonikolis, "Predictive discrimination and classification with mixed binary and continuous variables," *Biometrika*, vol. 77, no.3, pp. 657–662, 1990. doi:10.1093/biomet/77.3.657

[42] A. Willse, & R. J. Boik, "Identifiable finite mixtures of location models for clustering mixed-mode data," *Statistics and Computing*, vol. 9, no.2, pp. 111–121, 1999. doi:10.1023/A:1008842432747

[43] C. J. Lawrence, & W. J. Krzanowski, "Mixture separation for mixed-mode data," *Statistics and Computing*, vol. 6, no.1, pp. 85–92, 1996. doi:10.1007/BF00161577

[44] B. S. Everitt, "A finite mixture model for the clustering of mixed-mode data," *Statistics and Probability Letters*, vol. 6, pp. 305–309, 1988.

[45] W. G. Cochran, "Comparison of two methods of handling covariates in discriminatory analysis," *Annals of the Institute of Statistical Mathematics*, vol. 16, no.1, pp. 43–53, 1964. doi:10.1007/BF02868561

[46] J. H. Friedman, "Regularized Discriminant Analysis," *Journal of the American Statistical Association*, vol. 84, no.405, pp. 165, 1989. doi:10.2307/2289860

[47] M. M. B. Yvonne, E. F. Stephen, & W. H. Paul, "Discrete Multivariate Analysis Theory and Practice," New York, NY: Springer New York, 2007. doi:10.1007/978-0-387-72806-3

[48] U. Olsson, "On The Robustness Of Factor Analysis Against Crude Classification Of The Observations," *Multivariate Behavioral Research*, vol. 14, no. 4, pp. 485–500, 1979. doi:10.1207/s15327906mbr1404_7

[49] A. R. Leon, A. Soo & T. Williamson, "Classification with Discrete and Continuous Variables via General Mixed-Data Models," *Journal of Applied Statistics*, vol. 38, no.5, pp. 1021–1032, 2011.

[50] L. Amiri, M. Khazaei, & M. Ganjali, "General location model with factor analyzer covariance matrix structure and its applications," *Advances in Data Analysis and Classification*, vol. 11, no.3, pp. 593–609, 2017. doi:10.1007/s11634-016-0258-6

[51] P. Baah, A. Adebanji, & R. G. Kakaï, "Optimal ratio of continuous to categorical variables for the two-group location model," International *Journal of Applied Mathematics and Statistics*, vol. 42, no.12, pp. 18–26, 2013.

[52] E. Krusińska, R. Slowinski, & J. Stefanowski, "Discriminant versus rough sets approach to vague data analysis," *Applied Stochastic Models and Data Analysis*, vol. 8, no.1, pp. 43–56, 1992. doi:10.1002/asm.3150080107

[53] K. M. Lang, & W. Wu, "A Comparison of Methods for Creating Multiple Imputations of Nominal Variables," *Multivariate Behavioral Research*, vol. 52, no3, pp. 290–304, 2017. doi:10.1080/00273171.2017.1289360

[54] P. D. Allison, Missing Data. In SAGE Handbook of Quantitative Methods in Psychology (pp. 72–90). 1 Oliver's Yard, 55 City Road, London EC1Y 1SP United Kingdom: SAGE Publications Ltd, 2009. doi:10.4135/9780857020994.n4

[55] J. Franco, J. Crossa, J. Villaseñor, S. Taba, & S. A. Eberhart, "Classifying genetic resources by categorical and continuous variables," *Crop Science*, vol. 38, no.6, pp. 1688–1696, 1998. doi:10.2135/cropsci1998.0011183X003800060045x

[56] J. Franco, J. Crossa, J. Villaseñor, A. Castillo, S. Taba, & S. A. Eberhart, "A two-stage, three-way method for classifying genetic resources in multiple environments," *Crop Science*, vol. 39, no.1, pp. 259–267, 1999. doi:10.2135/cropsci1999.0011183X003900010040x

[57] L. Gutiérrez, J. Franco, J. Crossa, & T. Abadie, "Comparing a preliminary racial classification with a numerical classification of the maize landraces of Uruguay," *Crop Science*, vol. 43, no.2, pp. 718–727, 2003. doi:10.2135/cropsci2003.0718

[58] I. S. Andrade, C. A. F. Melo de, G. H. de S. Nunes, I. S. A. Holanda, L. C. Grangeiro, & R. X. Corrêa, "Morphoagronomic genetic diversity of Brazilian melon accessions based on fruit traits," *Scientia Horticulturae*, vol. 243, pp. 514–523, 2019. doi:https://doi.org/10.1016/j.scienta.2018.09.006

[59] B. P. Brasileiro, C. D. Marinho, P. M. A. Costa, L. A. Peternelli, M. D. V. Resende, D. E. Cursi, M. H. P. Barbosa, "Genetic diversity and coefficient of parentage between clones and sugarcane varieties in Brazil," *Genetics and Molecular Research*, vol. 13, no.4, pp. 9005–9018, 2014. doi:10.4238/2014.October.31.15

[60] R. N. F. Kurosawa, A. T. do Amaral Junior, F. H. L. Silva, A. D. dos Santos, M. Vivas, S. H. Kamphorst, & G. F. Pena, "Multivariate approach in popcorn genotypes using the Ward-MLM strategy: Morpho-agronomic analysis and incidence of Fusarium spp," *Genetics and Molecular Research*, vol. 16, no.1, 2017. doi:10.4238/gmr16019528

[61] R. Ortiz, J. Crossa, J. Franco, R. Sevilla, & J. Burgueño, "Classification of Peruvian highland maize races using plant traits," *Genetic Resources and Crop Evolution*, vol. 55, no.1, pp. 151–162, 2008. doi:10.1007/s10722-007-9224-7

[62] G. Padilla, M. E. Cartea, & A. Ordás, "Comparison of several clustering methods in grouping kale landraces," *Journal of the American Society for Horticultural Science*, vol. 132, no.3, pp. 387–395, 2007. doi:10.21273/jashs.132.3.387

[63] G. Padilla, M. E. Cartea, V. M. Rodríguez, & A. Ordás, "Genetic diversity in a germplasm collection of Brassica rapa subsp rapa L. from northwestern Spain," *Euphytica*, vol. 145, no.1-2, pp. 171–180, 2005. doi:10.1007/s10681-005-0895-x

[64] I. S. Andrade, C. A. F. de Melo, G. H. de Sousa Nunes, I. S. A. Holanda, L. C. Grangeiro, R. X. Corrêa, "Morphoagronomic genetic diversity of Brazilian melon accessions based on fruit traits," *Scientia Horticulturae*, vol. 243, pp. 514-523, 2019.

[65] H. Hamid, L. M. Mei, & S. S. S. Yahaya, "New discrimination procedure of location model for handling large categorical variables," *Sains Malaysiana*, vol. 46, no.6, pp. 1001–1010, 2017. doi:10.17576/jsm-2017-4606-20

[66] J. J. Daudin, "Selection of Variables in Mixed-Variable Discriminant Analysis," *Biometrics*, vol. 42, no.3, pp. 473–481, 1986.

[67] I. G. Vlachonikolis, & F. H. C. Marriott, "Discrimination with mixed binary and continuous data," *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, vol. 31, no.1, pp. 23–31, 1982.

[68] H. Hamid, P. A. H. Ngu, & F. M. Alipiah, "New smoothed location models integrated with PCA and two types of MCA for handling large number of mixed continuous and binary variables," *Pertanika Journal of Science and Technology*, vol. 26, no.1, pp. 247–260, 2018.

[69] H. Hamid, "New location model based on automatic trimming and smoothing approaches," *Journal of Computational and Theoretical Nanoscience*, vol. 15, no.2, pp. 493–499, 2018a. doi:10.1166/jctn.2018.7148

[70] H. Hamid, "Winsorized and smoothed estimation of the location model in mixed variables discrimination," *Applied Mathematics and Information Sciences*, vol. 12, no.1, pp. 133–138, 2018b. doi:10.18576/amis/120112