Systematic Literature Review of Mixed Variables Classification

Penny Ngu Ai Huong[1] and Hashibah Hamid[2]
[1]Awang Had Salleh Graduate School of Arts and Sciences, Universiti Utara Malaysia, 06010 UUM Sintok, Kedah, Malaysia
[2]School of Quantitative Sciences, Universiti Utara Malaysia, 06010 UUM Sintok, Kedah, Malaysia
[1]pennyngu90@hotmail.com [2]hashibah@uum.edu.my

Corresponding author email: pennyngu90@hotmail.com

***Abstract*— Classification is one of the most popular approaches that had been used in a variety of fields. There are a lot of classification methods that are applicable to classify objects into their respective groups. Among the classification methods, the location model and smoothed location model received attention when the data consists of mixed continuous and categorical variables. In this paper, classification, location model and smoothed location model are screened from the Scopus and Science Direct websites for review purposes. A total of 70 articles were chosen through the systematic review processes based on the related studies' topics. The studies had been reviewed and discussed under the topic of classification, location model and smoothed location model in different data situations. The systematic literature review has many benefits compared to traditional literature review. The systematic review process is able to provide a defined review process and fundamental priorities that can manage research bias easily. Based on the reviews, smoothed location model outperforms the other classification techniques in terms of classification performance. However, there are not many articles that particularly discuss the application of the location model and smoothed location model in mixed variables classification. Therefore, smoothed location model is suggested to be considered in the future work for mixed variables classification situation.**

Keywords—classification; location model; smoothed location model; systematic literature review

1. Introduction

The classification of mixed continuous and discrete variables has received a lot of attention over the past couple of decades. The development of statistical inference methods for evaluating mixed data is very important due to its applications that have been widely used in a variety of research fields including medical, physiological and social sciences. Moreover, experimental data are frequently gathered to predict the value of one or more system properties (responses), which can take either a quantitative or qualitative form. For example, in order to allocate the patients into one of the diagnostic or prognostic groups, plenty of diagnostic investigations in medicine involved the date collection made up of a combination of discrete and continuous factors (e.g., healthy and sick, malignant and benign, bad and good prognosis).

Various methods are available nowadays to analyze random variables that are comprised of both continuous and discrete variables. For instance, classification methods are used to examine the genetic diversity of accessions in gene banks by classifying accessions into subpopulations (subgroups or clusters) based on traits such as plant morphology, agronomic performance, disease resistance, genetic markers, and other factors [1]. There are many interesting problems that can be expressed where the response to be predicted is qualitative, as it may assume only a set of discrete values in the omics field [2]. An example would be the ability to distinguish between healthy and unhealthy individuals based on the experimental data gathered. In this situation, the two discrete values (groups) "sick" and "healthy" would represent the qualitative response to be anticipated, respectively.

Based on the experimental data gathered, classification methods are the suitable tools used to construct models with the goal of predicting which group most properly reflects the persons under investigation [2]. Software defect prediction is a quality control procedure that uses previous defect data along with software parameters. Prior to the software testing phase, the prediction technique can identify which software modules are defect prone. The majority of research studies employ machine learning classification techniques to divide software modules into two categories, which are defect-prone or non-defect-prone [3]. In the study of Kaya, et al [4], they applied different classification methods such as AdaBoostM1, Linear Discriminant, Linear Support Vector Machine (SVM), Random Forest, Subspace Discriminant, and Weighted-Knn (W-Knn) in their case studies as these methods are widely used in defect prediction studies.

In general, discriminant analysis is one of the data analysis that typically used to separate groups (categorical dependent variables) that are recognized as a priori, and their independent variables are quantitative variables and normally distributed [5]. One well-known model for assessing mixed data is the general location model (GLOM). The application of the location model is a famous technique in discriminant analysis when the independent variables employed include both qualitative (discrete) and quantitative (continuous) data. The location model is initially proposed by Olkin and Tate [6]. Given that the discrete variables are multivariate, normally distributed, with a constant covariance matrix across all cells indicated by the discrete variables, the location model assumes the conditional distribution of the continuous variables. [7]. Chang and Afifi [8] expanded the location model's application to two-group scenarios by creating a Bayes classification method for categorizing observations with both continuous and dichotomous variables. Krzanowski [9] had considered their results by generalization in which optimum and estimated

allocation rules were derived for mixed binary and continuous variables using likelihood ratio. Significant advancements have been made in the areas of expected error rate calculation, variable selection, heteroscedasticity of between-population dispersion, and heteroscedasticity of cross-location dispersion [10-13]. Krzanowski [11] provides an overview of the advances resulting from the location model and their possibilities for the future. In another article, two populations are categorized by using mixed covariates as well as discrete and continuous variables. It is assumed that the conditional dispersion matrices between the two populations are homogeneous and particular to the discrete values [14].

Limiting the number of discrete variables is necessary when using the location model to discriminatory situations in order to avoid having an excessive number of parameters that need to be estimated. If the initial sample sizes from each group are not big, Krzanowski [15] proposed the use of maximum six binary variables, with a corresponding decrease in number when some variables contain more than two states. The computing effort required to estimate misclassification rates were found to increase extremely with the number of binary variables, which is problematic if the initial sample sizes are big. Therefore, Krzanowski [15] suggested a discrete variable selection technique with backward elimination that can be utilized to find a suitable, reduced location model for discriminant applications where the number of binary variables is too large.

On the other hand, traditional maximum likelihood estimation in the location model encounters difficulties when empty cells exist due to many binary variables. To overcome this problem, a non-parametric smoothing method is proposed for parameters estimation in the location model [16]. Another case was the over-parameterization and instability of the covariance matrix in the location model, which was addressed by some authors by combining non-parametric smoothing and regularization [7]. More recently, Hamid [18] put up a concept that combines a principal component analysis technique for dimensionality reduction with a discriminant function based on the location model. The study's goal is to provide a different classification strategy when the observed variables are mixed and excessively huge.

A large number of current studies related to classification methods are carried out around the globe. As mentioned above, the use of mixed variables in classification is very common in different fields. Therefore, in this paper, a systematic review of the classification method is presented to ensure more exposure to the advantages of the classification method in different fields. Besides, not many studies that had been carried out for location model and smoothed location model for the purpose of mixed variables classification. Hence, the following section will perform the reviews on the location model and smoothed location model with mixed variables classification.

2. Materials and Methods

This part should contain sufficient detail so that all procedures can be repeated. It can be divided into subsections if several methods are described.

2.1 Identification

The systematic review process consists of three main stages that were used to select several relevant papers for this report. The first step entails the identification of keywords and the search for associated, related terms using a thesaurus, dictionaries, encyclopedias and prior research. Following the selection of all pertinent keywords, search strings on the Scopus and Science Direct databases (see Table 1) have been developed. The current research project was able to successfully retrieve 152 papers from both databases during the first stage of the systematic review process.

Table 1. The search string

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

During the initial screening phase, duplicate articles should be disregarded. The second phase screened 136 articles based on a number of inclusion-and-exclusion criteria created by researchers, while the first phase excluded 16 articles. Because literature (research articles) is the main source of useful knowledge, it was the first criterion. Additionally, publications in the form of systematic reviews, reviews, meta-analyses, meta-synthesis, book series, books, chapters, and conference proceedings are excluded from the current study. Additionally, the review was limited to English-language studies only 45 articles in all were disregarded based on particular criteria.

2.3 Eligibility

The third level, called eligibility, has a total of 91 articles ready. All article titles and significant content were now thoroughly inspected to make sure they matched the inclusion requirements and the goals of the current study. 21 articles were thus excluded because they are not related to either one of the topics. 70 articles are available for review in total (see Table 2).

Table 2. The selection criterion is searching

Criterion	Inclusion	Exclusion
Language	English	Non-English
Literature type	Journal (only research articles)	Journal (book chapter, conference proceeding)

2.4 Data Abstraction and Analysis

In this study, an integrative analysis was carried out in order to analyze and synthezise several research designs which include of qualitative, quantitative and mixed methodologies as depicted in Figure 1.

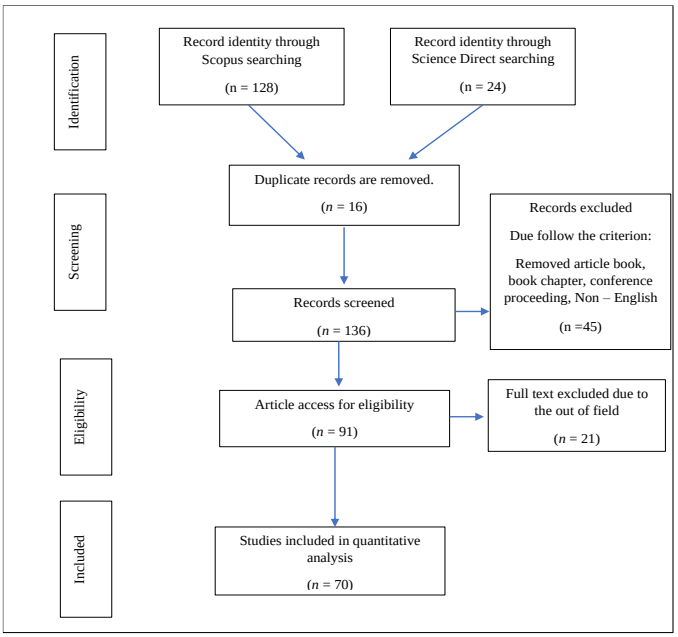

Figure 1. Flow diagram of the proposed searching study [73]

3. Results and Findings

In our real world, classification can be found anywhere whether in the field of medicine or manufacturing and many more. Based on the searching techniques, there are 70 articles that were extracted and analysed.

All articles were classified based on three main topics, which are classification (36 articles), location model (32 articles), and smoothed location model (2 articles) (refer Appendix Table 3).

3.1 Classification

One of the most significant current discussions is classification. Classification is a process of assigning each entity to a single class among a set of classes that are mutually exclusive and do not overlap. In the theory of library and information science (LIS), the classification is used to refer to three different but related concepts. The first concept is referred to a system of classes that are used to classify and organize a variety of activities in accordance with a predetermined set of guiding principles while the second concept is a class or group within a categorization system. Last but not least, classification is an act of putting objects into classes in a classification system [19].

There is a large volume of published studies describing the use of classification. In agriculture, classification also has been applied to categorize maize races or common beans. One of the studies categorize eight Peruvian highland maize races by using a numerical classification technique based on six vegetative features assessed over two years, and to compare this classification to the current racial classification [20]. The other study organizes common bean with Phaseolus vulgaris L. species accessions from various collection sites into groups based on how morphologically similar they were. The collection of bean accessions should be better maintained and used as a rich source of desirable features for plant breeding as a result of classification based on genetic diversity [21]. In another different study, genetic parameters and the Restricted Maximum Likelihood / best linear unbiased prediction (REML/BLUP) method are evaluated to predict Vitis genetic values [22]. Three homogeneous groups have been formed to identify the Vitis genotype groups based on the hybrids' resistance and broad sense heritability values. Another classification method, Regression analysis was developed to forecast the durability of structural wood members in a fire [23]. The durability of structural wood members in a fire depends on the member's cross-section brought on by the charring of the wood forms. The charring rate of the wood is determined by the char contraction factor, density, and moisture content and the result demonstrated the significance of moisture content and surface recession on wood charring.

Besides agriculture, numerous studies have attempted to explain the use of classification in classifying the species of animals too. For example, classification and regression trees (CART) and boosted regression trees (BRT) are the species distribution modeling (SDM) methods that are used to reevaluate the environmental parameters that influence coral reef distribution globally [24]. In another research, logistic regression (LR) and classification tree (CT) models are used to estimate microhabitat use and the summer distribution of juvenile Atlantic salmon, Salmo salar. The models predicted the presence or absence of salmon at that specific location based on some of the habitat variables (depth, current velocity, presence of instream and overhead cover, substratum particle size, and distance to stream bank) that were observed at that precise location [25]. In order to predict the species rarity of reptiles and amphibians in southern California, data on species presence and absence from 420 locations were used to create

classification trees, generalized additive models, and generalized linear models. Sensitivity, specificity, and the area under the curve (AUC) of the receiver-operating characteristic (ROC) plot based on twofold cross-validation, or bootstrapping, were used to assess the performance of the model. Climate, geography, soil, and vegetation characteristics were all predictors. The result showed that more precise species distribution models for rarer species were created by using a variety of modeling techniques [26].

Several researchers have reported that there has been an increasing amount of literature on the use of classification methods for different analysis purposes. Examples of these classification methods include Support Vector Machine, K-Nearest Neighbors, Naive Bayes algorithms, Regression, Naïve Bayes, Cluster Analysis and Convolutional Neural Networks. For instance, classification methods such as support Vector Machine, K-Nearest Neighbors, and Naive Bayes algorithms to analyze Twitter data to detect earthquakes as rapidly as feasible. Following the detection of a disaster, tweets are examined to identify persons connected to the event and demand demands and volunteer offers are gathered [27]. Another study used Support Vector Machine and bagged decision tree regression to categorize the activities of Bins and each pair of bins was employed as a feature and to predict the net metabolic cost. [28]. Besides that, two supervised classificatory approaches, maximum likelihood classification (MLC) and support vector machine (SVM) classification, are applied for image classification [29].

In 2021, Colaço and Abreu [30] introduced a new categorization scheme that depend on cluster analysis (k-means) and a limited number of variables density, diversity, and clustering is suggested where Lisbon had introduced this categorization in 1995, 2002 and 2010. The method can be generalized as a classification system, according to the cross-sectional examination of the commercial structures, which demonstrates how well it describes commercial location and change. The relationship between commercial classification and location model might be reinforced, strengthening the relevance of commercial studies in urban planning and policymaking. This is because the minimal dataset also permits the use of cluster membership on location models. This study offers some important insights that the cluster analysis is still able to attract the attention of researchers even in the year 2021. There is another study that used a classification for social media text streams during emergencies as the model is able to categorize social media text streams according to the topics they cover [31].

Distribution lines are essential to the modern power system and have a direct impact on the stability and security of the power supply. This is to prevent further financial and social expenses due to load interruptions. All occurrences of defects should be able to be quickly identified by a power system protection program. Fault diagnosis involves two tasks. One is fault classification, which has already attained excellent accuracy rates and the other one is fault location. There are two studies conducted to detect fault diagnosis by using the neural network and convolutional neural network (CNN). One of the research inspired by the Fourier transform provides an online data-driven method that converts signals from the time domain to the image domain using the signal-to-image (SIG) algorithm, and then processes the converted images using a convolutional neural network (CNN) architecture. The CNN-based structure, on the other hand, is significantly more compact than others. It would be simpler to transplant it to a hardware platform and would require less memory space [32]. Another study by Ferreira et al. (2020) suggested using autonomous neural networks to map the link between electrical signals at one terminal and transmission line defect information as this method gives an error range around the predicted short-circuit position [33].

In clinical diagnosis, 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 frequently used to detect the large biomolecules in trace amounts [34]. In one of the articles from the Journal of Microbiological Methods, a panel for the value of MALDI-TOF MS biomarker peaks and their correlation to outbreak strains, location, source, patient, diagnosis, and isolate genetics is constructed by using 55 well-characterized B. contaminants isolates [35]. Unsupervised clustering was implemented and classification models were generated using biostatistical analysis software. The result showed that MALDI-TOF MS may identify isolates that are not typeable by PFGE and successfully separates B. contaminants isolates into clonal, epidemiological clusters. Additional research should be done to better understand this ability.

3.2 **Location Model**

Linear discriminant function (LDF) that originated from Fisher [36] is always used in the case of discrimination and classification of new objects into one of two populations. However, some of the users restrict the usage of this method to only discrete variables and continuous variables. In 1975, Krzanowski [9] proposed a classification rule based on a location model that can deal with both binary and continuous variables. The proposed model is evaluated and applied in various data sets. The result showed that the classification performance based on the location model is able to give a satisfactory result compared to the original LDF. The author extended the research through further investigation of the location model by computing the Mahalanobis distances between these groups and visualizing the resulting configuration using principal coordinate analysis on breast cancer data in the year 1976. In 1977, Krzanowski [37] reviewed the performance of the LDF when some underlying assumptions of multivariate normality are violated. These assumptions are unequal variance-covariance matrices and non-normality data. The paper indicated that the location model is recommended for mixed binary and continuous data if LDF had poor classification performance. In 1980, Moussa [38] proposed a linear additive model to estimate parameters for the

location model when data consists of both binary and continuous variables with zero frequency or zero observations in some states.

Previous studies have reported discrimination between two populations when data comprises mixtures of binary and continuous variables. Therefore, in 1986, Krzanowski [39] extended the idea of discriminant rule from two populations to more than two populations. This proposed rule is then compared with the traditional normal-based rule. In 1993, Krzanowski [11] conducted a paper mainly to investigate the extensive developments and the capabilities of the location model by looking at the distance between groups, discriminant analysis, error rates, handling of missing data, their estimation and feature selection. In the following year, 1994, Krzanowski [12] further looked into the assumptions of the location model by relaxing the common within-cell dispersion matrices to allow the discrimination of different matrices between two populations. This change switched the Bayes location from a choice among linear functions to a chouse quadratic functions.

A number of researchers had extended the rule of location model. A predictive allocation rule for classification was derived and this rule is based on the usual frequency distribution of the location model and vague prior distribution for the unknown parameters [40]. The author compared the performance of the predictive rule and estimative rule in two cases, where binary variables provide no discrimination between two populations and there is discrimination between the population. The result stated that the predictive rule was able to give a lower misclassification rate for the first case while the estimative rule performed well in the second case. In another study, Willse and Boik [41] demonstrated that the finite mixture model by Lawrence and Krzanowski [42] cannot be identified without adding further limits. Therefore, the authors proposed the identifiable finite mixture models restricting the conditional means of the continuous variables. Simulations are used to evaluate these newly discovered models. The restricted location mixture models' conditional mean structure for the continuous variables is comparable to Everitt's [43] underlying variable mixture models, but the restricted location mixture models are easier to compute.

Leung [14] discussed the classification of mixed discrete and continuous variables in the presence of mixed covariates in the location model. Some of the variables are called covariates since they do not have obvious discrimination power in classification because of their same mean between the groups. These covariates are able to be used to produce a new variable for better classification [44]. A plug-in version of the Bayes rule with complete covariate adjustment is used to implement classification. Regularization in the general location model is proposed when the ratio of dimensionality of the continuous variables to the total training sample is less but close to unity [45]. Under certain conditions, a limiting overall expected error for the classifier is given and the error can be used to find the optimized regularisation parameters [13].

The general location model (GLOM) has been used in practice to treat the ordinal variables as nominal variables, while the conditional grouped continuous model (CGCM) is used to treat the nominal variables as ordinal during the discrimination analysis that involves mixed variables. However, these models might result in information loss when ordinal variables are categorized into nominal variables [46] and the latter model raise a concern about the model robustness when the ordinal variables are treated as continuous variables [47]. Therefore, a new proposed general mixed-data models (GMDMs) is proposed to treat the data that consist of mixtures of nominal, ordinal and continuous variables. This study proved that GMDMs is able to provide a satisfying classification performance in a study of croup in children [48]. Traditionally, in GLOM, it is assumed that continuous multivariate distributions across cells are generated by different categorical variable combinations with the same covariance matrices. In 2017, Amiri et al. [49] proposed a GLOM that takes into account both equal and unequal covariance matrices. The same factor analyser, factor analyser with unequal specific variance matrices (in the general and parsimonious forms), and factor analyser with shared factor loadings are the three covariance structures used across cells. These structures are used for both modeling covariance structure and for reducing the number of parameters. Three real data analyses serve as examples of the effective classification performance of these models.

Limiting the number of discrete variables is necessary when using the location model in discriminatory situations in order to avoid having an excessive number of estimated parameters that need to be estimated. The computing effort required to estimate misclassification rates were found to increase exponentially with the number of discrete variables, which is problematic if the initial sample sizes are big. Baah et al. [50] looked at a situation where the number of binary variables is a scalar multiple of the continuous variables as the parameters estimation in the location model is depend on the quantity of multinomial cells formed by the discrete variables and the continuous variables as well. The objective of the study was to identify a continuous-binary variable ratio combination that will provide the location model with the lowest possible error rates of misclassification for the two-group case.

Several studies investigating location models have been carried out by different researchers. For example, the performance of the location model approach from discriminant analysis is compared with a method that is based on rough sets theory in a set of real medical data [51]. Another study also compared the efficacy of five multiple imputations with chained equations (MI) techniques for handling imperfect nominal variables [52]. These methods are multiple imputations with chained equations (MICE), which uses polytomous regression as the elementary imputation method; classification and regression trees (CART); nested logistic regressions; Allison's [53] ranking procedure; and a joint modeling approach based on the general location model. Allison's [53] ranking method and MICE with CART fared badly in most conditions, but MICE with polytomous regression performed best.

The location model assumes that there are always observations in all possible combinations of the multinomial variable's values and the subpopulations. However, it's extremely likely that some of the multinomial cells may be empty in practical applications. Therefore, Franco et al., [54] proposed the modified location model (MLM) together with Ward's method in a two-stage of a clustering strategy. According to the results, the MLM may be used to analyse real data sets that contain empty cells. It seems suitable to use a two-stage strategy to locate primary groups using the Ward method and improve the composition of the groups using MLM. The authors further extended the research from two-way data to three-way data using the MLM with categorical and continuous variables. The strategy was then evaluated for classifying observations of one set of data simulation (with known structure) and two experimental data sets made up of multi-attribute, multi-site field trials of Caribbean and Conico accessions of Zea mays L maize [55]. The three-way Ward-MLM clustering strategy had also been applied in three different data sets with the goal of assessing the effectiveness of Ward-MLM methodology for grouping cultivars into low imperfect genotypic correlation (COI) groups [56].

Numerous studies have attempted to explain the use of Ward-MLM methodology in classifying gene bank accessions, genotypes or landraces. For example, Ward-MLM methodology is used for categorizing the gene bank accessions, classifying the Landraces into five cluster, identify redundant landraces, which allowed for a decrease in the number of accessions in subsequent crucial trials, grouping the 24 accessions of related Peruvian highland maize races and many more [57-62].

The location model faced the issue of empty cells because of the huge number of variables, especially binary variables. When there are too many variables and many empty cells, it could result in incorrect parameter estimates. This study suggests a technique for variable selection based on group distance, as determined by smoothed Kullback-Leibler divergence together with the location model [63]. Besides that, a concept is proposed by combining a principal component analysis-based dimensionality reduction technique with a discriminant function that based on the location model [64]. The study's goal is to provide practitioners with another viable tool for a classification problem that may be considered when the observed variables are mixed and excessively huge. Nevertheless, the author also suggested the nonlinear principal component analysis (NPCA) is incorporated into the classical location model (cLM), primarily to manage the high number of categorical variables, in order to reduce the high rate of misclassification. This investigation established the suggested model's new discrimination technique as a viable solution to the classification issues associated with mixed variables, particularly when dealing with large category variables [65].

3.3 **Smoothed Location Model**

The location model is one of the well-known methods that had been used for discriminant analysis in a multivariate data set that contains mixed categorical and continuous variables. However, the location model faced with overparameterization problem when there are many parameters involved in the analysis. This will lead to a situation where the number of binary variables is limited or a large number of training individuals is needed [11,66]. To solve this overparameterization problem, a study suggested non-parametric smoothing techniques when the range of applicability is significantly expanded while the number of parameters that must be estimated is drastically decreased [16]. The new proposed non-parametric smoothing techniques are compared with the other methods and it did provide good performance.

A useful technique that can handle both continuous and binary data simultaneously is the smoothed location model (SLM) [67]. However, SLM is not feasible when the data consists of great number of mixed variables. Hence, a variable extraction technique, principal component analysis (PCA) is integrated with the SLM to resolve the problem above [68]. Prior to constructing the smoothed location model, the principal component analysis was performed first and this method was evaluated on three actual data sets. The outcome of the study is satisfactory and the proposed method is recommended when handling with large number of mixed variables.

Besides PCA, in 2016 and 2018, to reduce the quantity of binary and continuous variables, another variable extraction method which is multiple correspondence analysis (MCA) combining with PCA had been conducted before the construction of the SLM [69-70]. Indeed, there are four types of MCA which are Indicator MCA, Burt MCA, Adjusted MCA and Joint Correspondence Analysis (JCA). This study aims to build a new SLMs by integrating SLM with two variable extraction methods which are PCA and two types of MCA. This is to reduce an excessive amount of mixed variables, mainly the binary variables. The performance of the new models developed is evaluated based on the misclassification rate, SLM+PCA+Indicator MCA and SLM+PCA+Burt MCA are evaluated (Hamid et al., 2018). The results show that the newly developed SLM models, when combined with two variable extraction methods, could be an effective alternative for addressing mixed variable problems for classification purposes, particularly when working with large binary variables.

Nonetheless, the location model is failed to perform when the data is contaminated with outliers [71]. A robust technique is implemented to accomplish the goal of reducing the impact of outliers in the construction of the discriminant rule based on the location model. In order to address the issues of outliers and empty cells simultaneously, a new location model is produced in the study. Winsorization is integrated with the smoothing method by considering the presence of outliers in the mixed variables. The findings have confirmed that the newly developed methodology and the new location model produced offer practitioner additional potential tools that may be taken into consideration in classification issues when the data samples contain outliers and that may also be used to address the

location model's crisis of some empty cells [72]. The statistical findings demonstrated that the newly developed location model performs optimally even for data with outliers. Additionally, experimental findings have supported the usefulness and practicality of the suggested classification approach.

4. Discussion

In this systematic review, we screened 135 journal articles and selected 70 articles that included classification methods, location models and smoothed location models. Through the review publications, we gained an overview of different classification methods that are had been applied in different filed. For instance, the support vector machine is usually used for image classification while the neural network is used for fault diagnosis [29,32].

From the past studies that have been reviewed, the location model is one of the best methods in performing mixed variables classification. However, if the maximum likelihood for parameters estimation is unavailable due to the empty cells, smoothed location model is a good option for this problem [16]. As mentioned, in previous studies, empty cells are created when there is a large number of binary variables [11]. According to the studies that are reviewed, there are a number of ways that are applicable to solve the problem. For example, the integration of variable extraction or variable reduction techniques with the location model or smoothed location model had been presented in several studies to solve the large binary variables problem [69-70].

5. Conclusions

This systematic review has analyzed the literature on the different classification methods especially the location model and smoothed location model in mixed variables classification. Based on the studies, smoothed location model gives a good classification performance when compared with the other classification methods. Nevertheless, there are only a few articles that specifically addressed the use of the location model and smoothed location model in mixed variables classification. The limited studies in the smoothed location model restricted more reviews on this model. Therefore, a future publication should be considered the smoothed location model when the data are in mixed variables.

Funding Statement

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

References

[1] G. Eason, B. Noble, and I.N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529-551, April 1955. (*references*)

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68-73.

[3] I.S. Jacobs and C.P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G.T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271-35

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

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

[8] 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.

[9] 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

[10] 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

[11] 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

[12] 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

[13] 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

[14] 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

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

[16] 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

[17] 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.

[18] 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.

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

[20] 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

[21] 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.

[22] 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

[23] 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

[24] 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

[25] 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

[26] 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

[27] 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

[28] 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

[29] 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

[30] 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

[31] 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

[32] 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

[33] 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

[34] 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

[35] 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

[36] 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

[37] 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

[38] 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

[39] 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

[40] 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

[41] 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

[42] 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

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

[44] 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

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

[46] 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

[47] 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

[48] 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.

[49] 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

[50] 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.

[51] 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

[52] 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

[53] 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

[54] 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

[55] 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

[56] 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

[57] 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

[58] 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

[59] 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

[60] 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

[61] 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

[62] 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

[63] N. I. Mahat, W. J. Krzanowski, & A. Hernandez, "Variable selection in discriminant analysis based on the location model for mixed variables," Advances in Data Analysis and Classification, vol. 1, no.2, pp. 105–122, 2007. doi:10.1007/s11634-007-0009-9

[64] 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.

[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. Hashibah, & N. I. Mahat, "Using principal component analysis to extract mixed variables for smoothed location model," Far East Journal of Mathematical Sciences, vol. 80, no.1, pp. 33–54, 2013.

[69] H. Hamid, N. Aziz, & P. N. A. Huong, "Variable extractions using principal component analysis and multiple correspondence analysis for large number of mixed variables classification problems," Global Journal of Pure and Applied Mathematics, vol. 12, no.6, pp. 5027–5038, 2016.

[70] 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.

[71] 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

[72] 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

[73] D. Moher, A. Liberati, J. A. D. Tetzlaff, PRISMA 2009 Flow Diagram. The PRISMA Statement, 2009.