TELKOMNIKA Telecommunication, Computing, Electronics and Control
Vol. 19, No. 1, February 2021, pp. 155~162
ISSN: 1693-6930, accredited First Grade by Kemenristekdikti, Decree No: 21/E/KPT/2018
DOI: 10.12928/TELKOMNIKA.v19i1.18024

A hybrid naïve Bayes based on similarity measure to optimize the mixed-data classification

Fatima El Barakaz, Omar Boutkhoum, Abdelmajid El Moutaouakkil
Department of computing, Laroseri Laboratory, University Chouaib Doukkali, El Jadida, Morocco

Article Info

Article history:

Received Jun 30, 2020
Revised Sep 8, 2020
Accepted Sep 16, 2020

Keywords:

CSBS
Mixed data
Multi-classification
Naïve Bayes
Short text
Similarity-based

ABSTRACT

In this paper, a hybrid method has been introduced to improve the classification performance of naïve Bayes (NB) for the mixed dataset and multi-class problems. This proposed method relies on a similarity measure which is applied to portions that are not correctly classified by NB. Since the data contains a multi-valued short text with rare words that limit the NB performance, we have employed an adapted selective classifier based on similarities (CSBS) classifier to exceed the NB limitations and included the rare words in the computation. This action has been achieved by transforming the formula from the product of the probabilities of the categorical variable to its sum weighted by numerical variable. The proposed algorithm has been experimented on card payment transaction data that contains the label of transactions: the multi-valued short text and the transaction amount. Based on K-fold cross validation, the evaluation results confirm that the proposed method achieved better results in terms of precision, recall, and F-score compared to NB and CSBS classifiers separately. Besides, the fact of converting a product form to a sum gives more chance to rare words to optimize the text classification, which is another advantage of the proposed method.

This is an open access article under the CC BY-SA license.

Corresponding Author:

Fatima El Barakaz
Department of computing, Laroseri Laborator
Faculty of the Sciences
Chouaib Doukkali University
Jabran Khalil Jabran Avenue B.P 299-24000, El Jadida, Morocco
Email: el.barakaz.fatima@gmail.com

1. INTRODUCTION

In many cases, datasets consist of both numerical and categorical variables. Many classifiers, such as linear regression, support vector regression, and k-nearest neighbour (KNN) are well-defined and validated for the computation of numerical variables. For these algorithms, it is easier to establish the relations between a target and its predictors when both are numerical. However, the numerical operations are not applicable to categorical variables, except if it has been converted to numeric one using coding systems such as dummy coding, effects coding, or even contract coding [1, 2]. Another approach is based on similarity and dissimilarity measures between categorical and numerical variables, where the data matrix is transformed into a distance configuration matrix after applying similar or dissimilar functions [3-5].

However, the previous approaches increase the number of predictors when categorical variables are numerous. In this case, the coding systems proposed additional steps to reduce the number of predictors [6, 7]. Though those approaches do not apply to multi-valued categorical variables that contain more than a single word, Mikolov proposes the Word2Vec model that represents the text in a vector format and saves

the syntax and the semantic meaning of natural language [8, 9]. The Word2vect is applicable even for a disordered multi-word text, where linguistic and semantic rules are not respected.

In the pre-processing and classification context, some approaches relying on similarity measure classification are applying cosine and string similarity to measure the distance between vectors. Other approaches propose utterly hybrid classifiers depending on the similarity-based measure. In this context, SBC algorithm (similarity-based classifier) [10] and CSBS (selective classifier based on similarities) are two algorithms that combine the measures of equality, reliability, and density to classify vectors. Both classifiers show excellent performance in terms of text classification [11, 12].

On the other hand, naïve Bayes (NB) is still highly useful to classify the categorical and numerical variables [13], especially compare its performance with other classifiers. In general, identifying suitable similarity measures between categorical variables or between categorical and numerical variables is considered a complex challenge. To address this challenge, a hybrid NB model has been constructed using an adapted CSBS. Where, the categorical variable is a short text, and we apply tokenization and stop-words in the pre-processing phase. For classification, NB has been used to train our model that used only the categorical variable. And for the portions that are poorly explained by NB, the adapted CSBS intervened in the second phase to improve the classification by including numerical variable.

The organization of the paper is as follows. Section 2 briefly presents the related works we address in the paper. Section 3 provides different methods used in this study. Section 4 introduces a description of the proposed hybrid naïve Bayes algorithm. Section 5 shows the experimental results of applying algorithms on the real credit card dataset. The last section presents the concluding remarks

2. LITERATURE REVIEW

2.1. Categorical variable and similarity measures

Categorical and qualitative multi-valued data have been studied for a long time in different contexts. Computing similarity has a long history, started with chi-square in the late 1800s that is frequently used for independence tests between categorical variables. Also, Pearson's chi-square has known many improvements that handled several data similarity cases [14]. So far, classical categorical data has changed. Notably, the categories number of a qualitative variable has increased to important values. Also, the categorical variables start to include multi-valued short text [10], so many limitations are exposed. Fortunately, different methods based on similarity measures have been proposed to overcome this challenge. However, the performance of those methods depends largely on data characteristics [15].

For the main data characteristics, we consider a categorical data contains N objects, with p categorical variables. While A_k denotes the k^{th}variable, and Ω_k the set of different values in A_k and n_k its cardinality. The key characteristics are the following:

- $f_k(x)$: The number of times the attribute $A_k$to take x as a value in a data set.
- $p_k(x)$: The sample probability of $A_k$to take x as a value in a data set, and it is given by;

$$p_k(x) = \frac{f_k(x)}{N} \quad (1)$$

- $p_k^2(x)$: Another probability formula of A_k to take x as a value in the given data set, and it's given by;

$$p_k^2(x) = \frac{f_k(x)(f_k(x)-1)}{N(N-1)} \quad (2)$$

In general, to measure a similarity value between two data instances X and Y belonging to a data set, all used measures respect the following form:

$$S(X,Y) = \sum_{k=1}^{d} w_k \, S_k(X_k, Y_k) \quad (3)$$

$S_k(X_k, Y_k)$: The per-attribute similarity between two values for the categorical attribute A_k.
w_k : The weight assigned to the attribute A_k, thereafter, it is fixed to 1/p.
The above expression has been the point of many studies and is interpreted into different functions depending on the data. Where three examples of $S_k(X_k, Y_k)$ and w_k have been mentioned. Starting with the sample one, the overlap measure: it counts the number of attributes that match in the two data instances, using the measure (4):

$$S_k(X_k, Y_k) = \begin{cases} 1 & if X_k = Y_k \\ 0 & if X_k \neq Y_k \end{cases} \quad (4)$$

The Goodall 4: measure: aims to normalize the similarity between two objects, based on the probability where the similarity value observed could be generated from a random sample of two points [16].

$$S_k(X_k, Y_k) = \begin{cases} p_k^2(x) & if X_k = Y_k \\ 0 & if X_k \neq Y_k \end{cases} \quad (5)$$

2.2. Bank customer transactions classification

Customer classification and targeting are widely applied in practice. In recent years, banks have invested in their data and applied machine learning methods for customer identification, where they achieved fruitful results. Eskin *et al*. [17] propose the use of a random sampling method to improve the support vector machine (SVM) model, for bank customer churn prediction. In the same context, De Caigny *et al.* [18, 19] suggested a combination of both methods of logistic regression and decision trees. While for fraud detection, Jurgovsky *et al.* showed how using long short-term memory (LSTM) improves the detection accuracy used the Random Forest classifier and incorporated transaction sequences [20]. Others focus on the pre-processing part, for the credit applications where various information about payment appear in qualitative, categorical attributes. In general, the classification of customer transactions could be used to extend a system that can compute socioecological impact from categorized transactions, and provide more analysis about the community and its relationship with the geographic location. And it is used in risk management, security and fraud detection, or commercial departments bank to identify customer behaviour.

2.3. Text classification

Text classification is a fundamental task in natural language processing. It is widely applied in sentiment analysis, recommendation and Fraud and spam detection [21, 22]. Machine learning includes many approaches for text classification as NB, support vector machine, and other algorithms. Lately, deep learning has shown an over-performing compared to traditional machine learning methods. And that is noticed in the known methods below: convolutional neural networks (CNNs) [23], recurrent neural networks (RNNs), and the combination of CNNs and RNNs [24].

Although the great success has shown in processing long sentences, it was not the case for short text explained by the data sparsity problem. Recently, many works have been applying various text presentation models to extract more information from short text [25, 26]. As mentioned earlier, some are based on features from multiple aspects, and others are based on transforming words into vectors. However, the text representations still face the data sparsity problem when the data include many new and rare words [27]. In our case, the text in question is categorized as a short text, where the variable is very multi-valued. So, the new and rare words cause a serious classification problem. In this paper, we propose a hybrid NB classifier based on adapted similarity measures applied to card transaction payment data.

3. RESEARCH METHOD

3.1. Naïve Bayes classifier

Naive Bayes is a supervised learning algorithm based on a probabilistic classification. This classifier is extremely faster compared to other methods. NB aims to calculate the joint probabilities of words and categories to estimate each category the text will be affected. The 'Naive' expression is due to the fact the words are independents. In other words, the conditional probability of a word from a category is assumed to be independent of the conditional probabilities of other words from the same category [28].

3.2. CSBS classifier

The CSBS is a classifier based on similarity measures, in which the treated limitations shown for short text classification are based on three measures: equality, reliability, and density [10]. For the sake of notation, for a class C, we distinguish between the amplitude a^c and the own amplitude a^{c*}, When the own amplitude of a given attribute serves to predict whether this is reliable relatively compare to other attributes, and that through eliminating the intervals containing values belonging to the other classes from a^c.

In CSBS classifier [11], equality is measured by the number of objects sharing the same values per attribute. The higher the measure is, the more the values indicate the membership to the class. However, the own amplitude indicates the reliably of the attribute. At the same time, an instance is more likely to belong to a class when the attribute value is included in its own amplitude. While the density of the membership of an instance to a class C is measured using the (6):

$$\xi_{lC} = \frac{1}{N}\sum_{j=1}^{M} p_{jC} \times \frac{N_j^C + a_j^{C*} - d(x_{lj}, \overline{c_j^C})}{N_j + a_j^C + \varepsilon} \quad (6)$$

where: M is the number of attributes.

N : The number of instances.

p_{jC} : The coefficient of reliability on x_j to predict C.

N_j^C : The number of instances that take the value of processed instance on attribute j^{th} per C.

a_j^{C*} : The own amplitude of C per attribute j^{th}.
a_j^C : The simple amplitude of C per attribute j^{th}.
$\overline{c_j^C}$: The center of C per attribute j^{th}.
N_j : The number of instances that take the value of processed instance on attribute j^{th}.
ε : A very small positive value.
Finally, the class of a given instance is the one having the highest membership measure using (7):

$$Y_{Instance} = argmax_{c\ \epsilon\ classes\ group}\ \xi_{IC} \quad (7)$$

4. THE PROPOSED METHOD

The primary purpose of the proposed algorithm is to provide a new hybrid algorithm that performs better for mixed data. This algorithm combines the individual strengths of NB for text application and CSBS. It mitigates the disadvantages of the two methods knowing that the performance of NB moves down where the number of rare words goes up. Besides, it has numerous advantages that can be described as follows:

- By combining a probabilistic algorithm with an algorithm based on distance and density, the model eliminates the probabilistic property of the proposed method.
- The computation complexity is lower compared to NB model as the proposed classifier turned the product form into a sum form.
- The impact of rare words number can not be ignored since it becomes an optimizer of classification performance.
- The CSBS contains a normalized distance, which is better for numerical variables applications.
- Implementation is more simple and easier.

The communicated advantages could be noticed through the algorithm's description as shown in Figure 1. The process shows the main steps to exceed the constraint due to NB fail to classify a particular instance, and the combination with the adapted CSBS in a specific stage. To illustrate the logic of our proposed model, Figure 2 represents the dealing of different components at each level. The trials' number is based on the value of K. For each trial the NB classifies the text instances based on the occurrence of words and the probabilities of belonging. However, and due to the high number of rare words, the NB affects an important portion to the wrong class. By adding the weight of the numerical dimension, the adapted CSBS tries to make the classification better and promote the position of each word in the dataset.

***Pre-processing**:*
1: let X-matrix denotes the data with M predictors: X-matrix= $[x_1, x_2, \ldots x_M]$ where p predictors are categorical and M-p+1 are numerical, and the class label is y.
2: Clean and pre-process X'-matrix= $[x_1, x_2, \ldots x_p]$, by removing the stops words and building the bags of words.
Training
3: Obtain the train set instances.
4: Train the NB model with the training dataset,
Classification
5: Obtain the test dataset.
6: Calculate the probabilities of belonging of the instance, using only the categorical part of the dataset.
4: Select the X-matrix train set corresponding to unclassified instances from the step above, and apply the adapted CSBS according to formula (8).
***Evaluation**:*
7: The model evaluation based on the performance of both NB and the adapted CSBS results, using K-fold cross validation.

Figure 1. The proposed algorithm

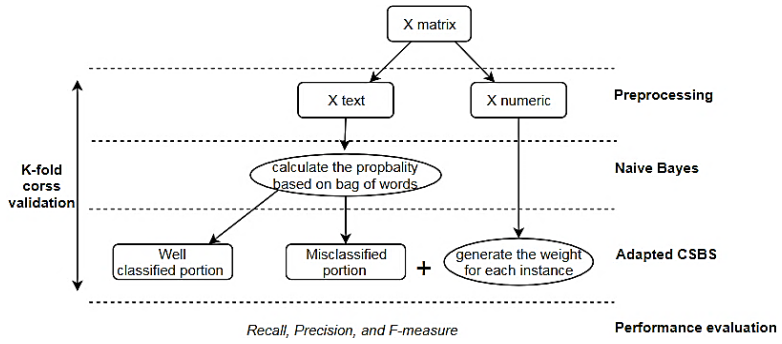

Figure 2. Illustration of different stages of proposed algorithm

5. RESULTS AND ANALYSIS

5.1. Experiments

5.1.1. Data description and preparation

The aim of our proposed solution is to effectively handle mixed data for card transactions payment classification problems. The dataset illustration contains 1312 instances and two variables. The first variable is a categorical variable that describes the transaction labels. The second is the numeric variable that consists of the amount associated with each operation. We extracted the data from a personal account created in Moroccan bank territory that we aim to classify them into four classes.

Observing our dataset, the categorical variable is an unstructured text and does not strictly respect the syntax or the semantic meaning of natural language (English, French...), or any abbreviation rules. Or either the emplacement of a word in a sentence does not have any importance. It could be categorized as a normal categorical dimension with few values, other cases contain multi-values, further, and it may also be classed as short text. In Table 1, each case has been presented with some selected instances.

The preparation of such data imposes three parts: tokenization, removal of stop words, then the construction of the bag of words. To tokenize the text of the categorical variable, strings of text have been split into words, we moved, and the stop words have been identified. For example: the, and, or... Stop words can also be a specified list of expressions, for example, taking the label: "Supermarket EL JADIDA", the expression "EL JADIDA" which is a name of a Moroccan city, has no sense in our proposed model, so our list of stop words combine the standard stop words in French and English languages list and the list of all Moroccan cities. Finally, the bag of words has been constructed as a matrix. This one helps the classifier to train on the data and recovers the significant terms of each class.

Table 1. Different cases selected from payment transaction text variable

Case	Payment transaction text	Comment
Standard Categorical dimension	"Achat YVES ROCHER MAROC" "Achat via WWW.ALIEXPRESS.COM" "Pay UBER MAROC E-COM bill"	Each instance belongs to different classes, and it appears in one form for the whole dataset.
Multi-value categorical dimension	"Achat Marjane market Alina" "Achat Marjane Bigdil" "Pay Marjane bill"	All instances belong to same classes, however the third one will be misclassified based using NB.
Short text	"Bill L'ARBRE DE ZOE" "Facture KINANI CHAUSSURES" "GRAS SAVOYE Molay Youssef"	The rare words are highly represented in this sample, the only keywords are "bill" and "facture", and the both are not enough to affect a correct classification with NB.

5.1.2. Experimental procedures

To evaluate the proposed algorithm, we train with three models. The first is NB, which was applied to the categorical variable to avoid the overlapping of the numerical variable. The second model used the adapted CSBS on both categorical and numerical variables. The last one introduced our proposed model that combines the NB and the adapted CSBS algorithm. To adjust the CSBS (cited in (6)) to the structure of the dataset. The adapted CSBS is given in the (8):

$$\xi_{lC}(\mathrm{X}) = \frac{N_t^C + a_t^{C*} - d(x, \overline{c_t^C})}{N_t + a_t^C + \varepsilon} \times \frac{1}{N} \sum_{j=1}^{M'} p_{jC} \quad (8)$$

where: p_{jC} indicate the frequency of the word w_j per class C.
t: used to index the parameters of the numerical attribute.
M′: Number of words of the categorical variable

For a reasonable comparison, we organized the dataset into different subset sizes, n=280, 560, 840, and 1120, respectively, which are selected each time arbitrarily from our dataset of 1312 instances. The K-Fold Cross-validation sampling method is frequently used to evaluate models in machine learning and data mining. The dataset is segmented randomly into K segments, where each segment is retained once, and the classifier is learned on the other K-1 segments. In our case, K will take 4, 7, and 10, respectively.
Therefore, the learning procedure is performed K times on each different subset. The overall performance is evaluated in terms of recall, precision, and F-measure:

$$\mathrm{Precision} = \frac{TP}{TP+FP} \quad (9)$$

$$\mathrm{Recall} = \frac{TP}{TP+FN} \quad (10)$$

$$\mathrm{F\_score} = 2 \times \frac{Recall \times Precision}{Recall + Precision} \quad (11)$$

where: FN is the number of false negatives.
FP is the number of false positives.
TP is the number of true positives.
The calculation of those two factors in a multi-class classifier situation request the notions below:
Classified

$$C = \text{Actual} \begin{matrix} c_{11} & \dots & c_{1n} \\ \dots & & \\ c_{n1} & & c_{nn} \end{matrix}$$

The confusion elements for each class are given by:

$$tp_i = c_{ii} \quad ; \quad fn_i = \sum_{l=1}^{n} c_{il} - tp_i \tag{12, 13}$$

$$fp_i = \sum_{l=1}^{n} c_{li} - tp_i \tag{14}$$

$$tn_i = \sum_{l=1}^{n} \sum_{k=1}^{n} c_{kl} - tp_i - fp_i - fn_i \tag{15}$$

5.2. Experiments results:

The performance evaluation of our hybrid model constructed using K-fold cross-validation introduced in the section above. Since the parameter K took different values, we compute the model on 30 trials for each sample size. The results for the three classifiers NB, adapted CSBS, and the proposed method are reported in Table 2. The improvements of the hybrid method in terms of the different measures refer at first to the performance of naïve Bayes on the dataset, then at second to the adding of the adapted CSBS performance applied to the partitions poorly classified. Furthermore, the notable role of the adapted CSBS could not be denied, since it kept an excellent harmonic mean between the recall and the precision for each different simulation. And better, when it is combined with NB performance. To present the progress of our classifier in term multi-classification improvement, we selected for K=10 four trials randomly applied on a sample of n=280. And based on Table 3, which describes the recall, precision, and F-score values, the proposed method outperformed for the three evaluation indicators.

Table 2. The results of the different classifier for different K value, based on 30 trials on average

		Naive Bayes			Adapted CSBS			The proposed model		
	Sample size	Recall	Precision	F-score	Recall	Precision	F-score	Recall	Precision	F-score
K=4	280	0.63	0.76	0.62	0.78	0.79	0.83	0.79	0.89	0.89
	560	0.61	0.73	0.62	0.75	0.82	0.79	0.78	0.89	0.86
	840	0.72	0.71	0.71	0.83	0.89	0.77	0.88	0.93	0.86
	1120	0.76	0.68	0.72	0.76	0.75	0.72	0.89	0.89	0.94
K=7	280	0.71	0.75	0.64	0.78	0.84	0.75	0.84	0.92	0.93
	560	0.78	0.69	0.62	0.84	0.74	0.64	0.8	0.88	0.85
	840	0.63	0.79	0.72	0.65	0.87	0.63	0.83	0.94	0.83
	1120	0.67	0.71	0.74	0.74	0.85	0.71	0.98	0.91	0.89
K=10	280	0.6	0.61	0.62	0.77	0.89	0.62	0.88	0.8	0.88
	560	0.7	0.6	0.62	0.83	0.81	0.73	0.84	0.97	0.8
	840	0.76	0.8	0.71	0.74	0.74	0.66	0.77	0.96	0.89
	1120	0.78	0.67	0.72	0.72	0.84	0.77	0.9	0.88	0.94

Table 3. The results of precision, recall, and F-score per trial and per method

		Method	Recall	Precision	F-Score
Trial.1	1	Naive Bayes	0.78	0.89	0.83
	2	Adapted CSBS	0.74	0.76	0.75
	3	Proposed method	0.89	0.94	0.91
Trial. 2	4	Naive Bayes	0.9	0.85	0.88
	5	Adapted CSBS	0.78	0.77	0.78
	6	Proposed method	0.94	0.91	0.92
Trial. 3	7	Naive Bayes	0.87	0.9	0.88
	8	Adapted CSBS	0.8	0.74	0.77
	9	Proposed method	0.93	0.94	0.94
Trial. 4	10	Naive Bayes	0.83	0.85	0.84
	11	Adapted CSBS	0.77	0.75	0.76
	12	Proposed method	0.9	0.93	0.91

Even more, the hybrid method guarantees a good efficiency in terms of the one class classification performance, so we have:

$$\text{Precision } (C_{NB}) < \text{Precision } (C_{\text{The poposed method}}) \quad \text{And:} \quad \text{Recall}(C_{NB}) < \text{Recall}(C_{\text{The poposed method}})$$

To visualize this, enhance, a demonstration with a confusion matrix is recommended. Figure 3 illustrates the confusion matrix of different selected trials per method. Moving from NB to adapted CSBS to the proposed method for each trial, the numbers in the confusion matrix increased where the numbers outside decreased, which proves the progress of one-class classification. We also note that the True Positive in tables (3), (6), (9), and (12) are better than its equivalent in tables (2), (5), (8), and (11). This result highlights the fact of how the hybrid method works significantly better for the rare words and achieved excellent results for both mixed data classification and text classification. In general, the NB shows good results comparing to the results of CSBS. However, the combination of both achieved meaningful classification progress.

1)	C1	C2	C3	C4
C1	27	2	7	0
C2	2	86	4	0
C3	1	2	103	1
C4	1	4	22	18

2)	C1	C2	C3	C4
C1	27	2	1	15
C2	2	70	3	1
C3	1	2	110	1
C4	1	9	17	18

3)	C1	C2	C3	C4
C1	37	0	2	2
C2	0	79	4	0
C3	1	2	115	0
C4	1	0	10	27

4)	C1	C2	C3	C4
C1	56	0	1	2
C2	2	40	4	1
C3	22	2	141	2
C4	1	3	0	3

5)	C1	C2	C3	C4
C1	49	2	2	15
C2	2	40	8	2
C3	8	3	141	0
C4	1	4	0	3

6)	C1	C2	C3	C4
C1	57	0	1	2
C2	7	48	0	1
C3	3	0	155	2
C4	0	1	0	3

7)	C1	C2	C3	C4
C1	40	9	7	0
C2	2	104	2	1
C3	5	0	68	2
C4	0	2	6	32

8)	C1	C2	C3	C4
C1	47	11	5	0
C2	2	82	4	0
C3	3	3	79	1
C4	1	6	4	32

9)	C1	C2	C3	C4
C1	58	8	0	0
C2	0	91	1	0
C3	2	2	82	1
C4	1	0	2	32

10)	C1	C2	C3	C4
C1	50	9	0	2
C2	11	61	10	0
C3	0	13	84	0
C4	2	1	5	32

11)	C1	C2	C3	C4
C1	52	3	4	2
C2	2	66	2	0
C3	2	13	88	7
C4	0	2	5	32

12)	C1	C2	C3	C4
C1	63	0	0	0
C2	2	69	10	0
C3	0	3	89	0
C4	0	0	5	39

Figure 3. The confusion matrices of four trials were randomly selected to explain the result of Table 3

6. CONCLUSION

The main objective of this contribution is to deal with the classification of mixed data that include a multi-valued short text variable. We introduced a hybrid naïve Bayes that is based on similarity measures to effectively process both categorical and numerical variables. In the proposed method, the naive Bayes predicts the portion of the target only explained by the categorical variable, and the remaining part is predicted using the adapted CSBS that provides good classification using numerical variables. The proposed solution combines NB with an adapted CSBS. The hybrid model was compared to the naïve Bayes, and the adapted CSBS separately. The experiments were performed using the card transactions payment data that contains a multi-valued short text variable and numerical variable. The solution has achieved significant progress in terms of recall, precision, and F-measure. Furthermore, it deals well with rare words issues, and also improves the classification of the model.

This work is limited because it has not been applied to different known dataset yet. However, it was proposed to handle the classification of short text using multi-valued variables, applied to a real case problem: card transaction payment classification. This study could be extended on many mixed datasets in a different field in order to optimize the classification of categorical dimensions. In future work, the dimensionality of vector-text supported by our method will be investigated while maintaining its simplicity.

ACKNOWLEDGEMENTS

This study was supported by the Research team at INDATACORE, a company of artificial intelligence solutions.

REFERENCES

[1] Cohen, "Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences," *Amazon Warehouse, Fulfilled by Amazon*, 2013.

[2] C. M. Cuadras, C. Areans, and J. Fortiana, "Some computational aspects of a distance—based model for prediction," *Communications in Statistics - Simulation and Computation*, vol. 25, no. 3, pp. 593-609, 1996.

[3] C. Cuadras and C. Arenas, "A distance-based regression model for prediction with mixed data," *Communications in Statistics - Theory and Methods*, vol. 19, no. 6, pp. 2261-2279, 1990.
[4] E. B. D. Val, M. M. C. Bielsa, and J. Fortiana, "Selection of Predictors in Distance-Based Regression," *Communications in Statistics - Simulation and Computation*, vol. 36, no. 1, pp. 87-98, 2007.
[5] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 1, pp. 49-67, 2006.
[6] L. Meier, S. V. D. Geer, and P. Bühlmann, "The group lasso for logistic regression," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 70, no. 1, pp. 53-71, 2008.
[7] V. K. Ayyadevara, "Word2vec," *Pro Machine Learning Algorithms*, pp. 167–178, 2018.
[8] A. Neelakantan, J. Shankar, A. Passos, and A. Mccallum, "Efficient Non-parametric Estimation of Multiple Embeddings per Word in Vector Space," *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.
[9] P. Venkateswari, P. Umamaheswari, K. Rajesh, J. Glory Thephoral, "Gene based Disease Prediction using Pattern Similarity based Classification," *International Journal of Innovative Technology and Exploring Engineering Regular Issue*, vol. 8, no. 11, pp. 3223-3227, 2019.
[10] A. Skabar, "Direction-of-Change Financial Time Series Forecasting using a Similarity-Based Classification Model," *Journal of Forecasting*, vol. 32, no. 5, pp. 409-422, 2013.
[11] W. Cherif, A. Madani, and M. Kissi, "A Novel Similarity-Based Algorithm for Supervised Binary Classification: Sandalwood Odor Application," *SSRN Electronic Journal*, 2018.
[12] S. Chen, G. I. Webb, L. Liu, and X. Ma, "A novel selective naïve Bayes algorithm," *Knowledge-Based Systems*, vol. 192, 2020.
[13] Z. E. Rasjid and R. Setiawan, "Performance Comparison and Optimization of Text Document Classification using k-NN and Naïve Bayes Classification Techniques," *Procedia Computer Science*, vol. 116, pp. 107-112, 2017.
[14] B. C. Brookes and H. Cramer, "The Elements of Probability Theory and Some of Its Applications," *The Mathematical Gazette*, vol. 40, no. 332, p. 153, 1956.
[15] Z. Šulc and H. Řezanková, "Evaluation of Recent Similarity Measures for Categorical Data," *International Scientific Conference*, 2014.
[16] D. W. Goodall, "A New Similarity Index Based on Probability," *Biometrics*, vol. 22, no. 4, pp. 882-907, 1966.
[17] E. Eskin, A. Arnold, M. Prerau, L. Portnoy, and S. Stolfo, "A Geometric Framework for Unsupervised Anomaly Detection," *Advances in Information Security Applications of Data Mining in Computer Security*, pp. 77-101, 2002.
[18] A. D. Caigny, K. Coussement, and K. W. D. Bock, "A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees," *European Journal of Operational Research*, vol. 269, no. 2, pp. 760–772, 2018.
[19] G. Nie, W. Rowe, L. Zhang, Y. Tian, and Y. Shi, "Credit card churn forecasting by logistic regression and decision tree," *Expert Systems with Applications*, vol. 38, no. 12, pp. 15273-15285, 2011.
[20] J. Jurgovsky, M. Granitzer, K. Ziegler, S. Calabretto, P.-E. Portier, L. He-Guelton, and O. Caelen, "Sequence classification for credit-card fraud detection," *Expert Systems with Applications*, vol. 100, pp. 234-245, 2018.
[21] Y. Su, Y. Huang, and C.-C. J. Kuo, "Efficient Text Classification Using Tree-structured Multi-linear Principal Component Analysis," *arXiv.org*, 24-Feb-2018.
[22] H. Kauderer and H.-J. Mucha, "Supervised Learning with Qualitative and Mixed Attributes," *Classification, Data Analysis, and Data Highways Studies in Classification, Data Analysis, and Knowledge Organization*, pp. 374-382, 1998.
[23] H. Chen, M. Sun, C. Tu, Y. Lin, and Z. Liu, "Neural Sentiment Classification with User and Product Attention," *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016.
[24] Y. Su, Y. Huang, and C.-C. J. Kuo, "Efficient Text Classification Using Tree-structured Multi-linear Principal Component Analysis," *arXiv.org*, 24-Feb-2018.
[25] W. Hua, Z. Wang, H. Wang, K. Zheng, and X. Zhou, "Short text understanding through lexical-semantic analysis," *2015 IEEE 31st International Conference on Data Engineering*, pp. 495–506, 2015.
[26] Y. Su, R. Lin, and C.-C. J. Kuo, "Tree-structured multi-stage principal component analysis (TMPCA): theory and applications," arXiv.org, 2018.
[27] R. Malik, "Learning a classification model for segmentation," *Proceedings Ninth IEEE International Conference on Computer Vision*, 2003.
[28] N. Sharma, M. Singh, "Modifying Naive Bayes classifier for multinomial text classification," *2016 International Conference on Recent Advances and Innovations in Engineering (ICRAIE)*, pp. 1-7, 2016.