A Novel Feature Selection Method Based on Clustering

Jonathan A. Mata-Torres
Reynosa-RODHE Multidisciplinary Unit
Autonomous University of Tamaulipas
Reynosa, Tamaulipas, México
e-mail: a2093010058@alumnos.uat.edu.mx

Edgar Tello-Leal
Faculty of Engineer and Science
Autonomous University of Tamaulipas
Victoria, Tamaulipas, México
e-mail: etello@uat.edu.mx

Uluses M. Ramirez-Alcocer
Faculty of Engineer and Science
Autonomous University of Tamaulipas
Victoria, Tamaulipas, México
e-mail: a2093010066@alumnos.uat.edu.mx

Gerardo Romero-Galván
Reynosa-RODHE Multidisciplinary Unit
Autonomous University of Tamaulipas
Victoria, Tamaulipas, México
e-mail: gromero@docentes.uat.edu.mx

Abstract— **Nowadays, there is a great interest from academia, the industry, and the government to find potentially useful information to build a prediction model from data with high dimensionality, which has become one of the most important challenges in data mining and machine learning approaches. In this way, feature selection is the process of selecting the most useful features for building models in tasks like classification, regression or clustering, in order to reduce the dimensionality and facilitating the visualization and understanding of the data. In this paper, we propose a feature selection method based on the mean shift clustering algorithm and the Pearson correlation coefficient to contribute to solving some of the challenges in the data analytics systems, of real-time execution. Furthermore, we compare the mean shift method with the renowned Recursive Feature Elimination (RFE) method, as well as with the feature selection method designed by a human expert in the domain. Finally, the subsets of data generated with the attributes selected by the methods are evaluated by the J48 classification algorithm based on a decision tree, using a historical public safety data set. The clustering method proposed has a great advantage over the other methods in the computing time required to recommend a group of selected attributes.**

Keywords–feature selection; mean shift; clustering; data mining; J48.

I. Introduction

Nowadays, the trend in the administration of resources, infrastructure, and services in cities is increasingly based on their ability to make decisions using knowledge bases, as well as their potential to anchor external knowledge and the implementation of knowledge-based strategies, in order to provide a better quality of life to the citizens and visitors. This way, the concept of smart cities emerged, in which a smart city can understand how an urban environment is capable of offering advanced and innovative services for citizens in order to improve the quality of life in general by using widespread support of systems (system of systems) based on Information and Communication Technologies (ICT) [1]-[3]. ICT software applications and the intensive use of digital devices such as sensors, actuators, and mobiles are essential means for realizing smartness in any of smart city domains [4]. A concept closely related to smart cities is the Internet of Things (IoT) [5], representing an extension of the Internet with a large number of objects (physical or virtual things) with pervasive sensing, detection, actuation, and computational capabilities allowing these devices to generate, exchange, and consume data with minimal human intervention [6][7]. In smart cities, specific areas of application have been identified through smart systems, e.g., transportation, public safety, sustainability, healthcare, energy, transportation and mobility, environment, education, and governance [8][9].

The automation of a large number of business processes and transactions that run on inter-organizational information systems within smart cities, embedded systems, smart systems based on IoT technology, as well as the intensive use of social networks through smartphones and software systems that use cloud computing technology, have caused the generation of massive volumes of data (known as big data), of different types: structured, semi-structured or unstructured [10][11]. The knowledge extraction and the hidden correlations of big data is a growing trend in information systems to provide better services to citizens and support decision-making processes [12].

There is a great interest from academia, the industry, and government for the development and deployment of big data analysis applications, both for general use and specific use in smart cities, which face different challenges. Hence, finding potentially useful information to build a prediction model from data with high dimensionality has become one of the most important challenges of data extraction and knowledge discovery [13][14]. One of the effects of the high dimensionality in the data sets can cause prediction models with a low precision measure. In addition, these is a high computational cost associated with processing of a big

Copyright (c) IARIA, 2019. ISBN: 978-1-61208-731-3

volume of data to predict an event [15]. To solve problems of high dimensionality in data sets (dimensionality reduction), approaches based on the feature selection and feature extraction methods have been proposed.

Feature selection methods are used in data mining and machine learning, commonly in the pre-processing stages, and include both supervised and unsupervised techniques [16]. A feature is an individual measurable property of the instance being observed, and, through a set of attributes, a data mining or machine learning algorithm can perform data classification or clustering [17]. The feature selection approaches aim to select a small subset of features that minimize redundancy and maximize relevance to the target such as the class labels in classification [18]. In other words, the feature selection consists of selecting a subset of features representative of the original data set, but that can efficiently describe the input data.

The feature selection algorithms can be classified into the following categories: filter, wrapper, and hybrid methods [15]. The filter methods select the most relevant features using variable ranking techniques as the principle criteria for attribute selection. In filter methods, the features weights are individually calculated based on some criteria (e.g., correlation coefficient), the attributes that satisfy these conditions are considered as selected features and the remaining ones are removed from the subset [19]. Furthermore, the wrapper methods use a predictor as a black box and the predictor performance as the objective function to evaluate the variable subset [17]. In wrapper methods, several search algorithms can be used to find a subset of variables which maximizes the objective function which is the classification performance. That is, in wrapper method uses the information of the classifier to find the best feature subset, usually by performing computationally expensive searches on the feature space. Additionally, the hybrid methods try to exploit the best functionalities of the filters and wrappers approach, trying to reduce the computational cost but maintaining the effectiveness in the objective task associated with using the selected functions [20].

In this paper, we propose a feature selection method based on the mean shift clustering algorithm in combination with the Pearson correlation measure, allowing to identify a subset of relevant and non-redundant attributes. In addition, we compare the mean shift method with the renowned Recursive Feature Elimination (RFE) method [21], as well as with a feature selection method designed by a human expert in the crime domain. Finally, the methods are evaluated through their implementation in a classification algorithm based on J48 decision tree. In the proposal, a data set of crime incidents from the last 17 years is used, where their records are collected by a set of software systems implemented in a smart city. The method based on the mean shift clustering algorithm has a great advantage over the other methods in the processing time required to recommend a group of attributes selected.

The rest of the paper is organized as follows. The feature selection algorithm is the subject of Section 2. The experimental study and results are covered in Section 3. The conclusions and future research are presented in Section 4.

II. Clustering feature selection method

The proposed feature selection algorithm integrates the concepts of the mean shift clustering algorithm and the Pearson correlation analysis. The former is an unsupervised learning algorithm that clusters the data based on its natural distribution. This algorithm is characterized by not requiring prior knowledge of the number and location of the centroids. In the other hand, the Pearson correlation measures the statistical relationship between two variables, specifically the dependence of one variable on another variable. Therefore, in a statistical correlation, the two variables that are correlated are dependent on each other and one may be used to predict the other. The mean shift algorithm is a statistical clustering method based on non-parametric kernel density estimation, which is expressed by (1). Given *n* data points $x_i,\ i = 1,\ldots, n$ in the d-dimensional space R^d, the multivariate kernel density estimator with kernel *K(x)* and a symmetric positive definite *d x d* bandwidth matrix **H**, computed in the point *x* is given by [22][23]:

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} K_H(x - x_i), \qquad (1)$$

In Fig. 1 shows the operation of the method. The input data set is represented by a numerically encoded matrix. This data set turn into a data set or transposed matrix, i.e., let A be a matrix of dimension *m x n*, we denote the element of row *i* and column *j* as *A(i, j),* where $i < m$ *and* $j < n$. Then, the transposed matrix of *A* is defined as the matrix A^T of dimension *n x m* such that $A^T(j, i) = A(i, j)$, where $i < m$ and $j < n$. Next, the mean shift algorithm initializes a window on all data points; with the first data point, its distance from all data points is calculated. The data points are used to find a new mean *m(x)* of the window according to the kernel *k(x)* [24]. The iterations continue until the mean of a window becomes fixed. Then, the algorithm will move on to the second data point and repeat the same procedure. The iterations will continue until the system converges.

The mean shift algorithm generates a list with the set of clusters, which contains all the data set objects. The elements of the list are compared to each other to determine if they belong to the same clusters. If that is the case, the Pearson correlation coefficient between the two objects *(i,j)* is calculated by (2), using the original data set.

$$R(i) = \frac{cov(x_i, Y)}{\sqrt{var(x_i) * var(Y)}}, \qquad (2)$$

where x_i is the i^{th} variable, *Y* is the output (class labels), *cov()* is the covariance and *var()* the variance. Correlation

Copyright (c) IARIA, 2019. ISBN: 978-1-61208-731-3

ranking can only detect linear dependencies between variable and target.

The calculation of the Pearson correlation generates a correlation matrix, to which a threshold (filter) of $\geq +0.5$ and -0.5 is applied, which allows us to remove the attributes with a value below the threshold, that is, with low linear correlation, allowing to select the representative attributes of the data set.

```
Algorithm 1 Feature Selection Method
Input: Dataset: Numeric encode Matrix
Output: FeatureSelected: list
    MSHIFTP(Dataset) :
    $Clusters \leftarrow MeanShift(Dataset^T)$
    for $i$ in range(0, len($Clusters$)) do
        for $j$ in range(0, len($Clusters$)) do
            if $Clusters(i) == Clusters(j)$ then
                $PearsonL.Add(Pearson(Dataset[i], Dataset[$
            end if
        end for
    end for
    for $x$ in range(0, len($PearsonL$)) do
        if $PearsonL[x] \leq -0.5$ or $PearsonL[x] \geq 0.5$ then
            $FeatureSelected.Add(PearsonL[x])$
        else
            $Discard()$
        end if
    end for
    return $FeatureSelected$
```

Figure 1. Feature selection algorithm.

III. EXPERIMENTAL STUDY

In this section, we compare the performance of our proposed clustering feature selection method with the RFE method and with a human expert method. A crime incidents data set of a smart city is used in our experiment. In this data set, crime data has been collected through a set of information systems and IoT technologies. The incidents of crime are from the City of Chicago, in the period from 2001 to 2017, consisting of a total of 6.4 million records and 22 attributes [25].

Table I shows a description of the attributes contained in the data set. These attributes can be values of data type: string, numeric, date, location or Boolean. Further, the total number of cases or values contained by attribute are shown. The ID, Case Number, and Date attributes are not used in the execution of the attribute selection method, and the arrest attribute represents the class label of the data set, of Boolean type.

A. Feature Selection Results

In the feature selection mean shift-based method, the total number of records contained in the data set (6.4 million) was used to make the recommendation of the attributes to be selected. This method selects 9 attributes (UICR, FBI Code, Y coordinate, Latitude, Location, Beat, District, Ward, and Community Area) from a total of 17.

TABLE I. DESCRIPTION OF THE ATRIBUTES CONTAINED IN THE DATA SET

Feature	*Description*	*No cases*
ID	Unique identifier for the record	6,457,411
Case Number	Unique identifier of the incident assigned by the Chicago policy department.	6,457,411
Date	Date when the incident occurred.	2,740,512
Block	Extract of the address where the incident occurred.	60.144
IUCR	Codes used to classify criminal incidents by law enforcement agencies.	350
Primary Type	The primary type description of the IUCR code.	35
Description	The secondary description of the IUCR code	380
Location Description	Description of the location where the incident occurred.	180
Domestic	Indicates whether the incident is related to violence domestic.	2
Beat	Indicates the police district where the incident occurred.	25
Ward	The ward (City council district) where the incident occurred.	50
Community Area	Indicates the community area where the incident occurred	77
FBI Code	Indicates the crime classification based in the FBI system	26
X coordinate	The X coordinate of the location where the incident occurred in the state of Illinois.	78,528
Y coordinate	The Y coordinate of the location where the incident occurred in the state of Illinois.	129,825
Year	Year when the incident happened	17
Update On	Date and time when the record was updated.	2,593
Latitude	The latitude of the location where the incident took place.	861,599
Longitude	The longitude of the location where the incident happened	861,046
Location	This attribute is formed with the data of latitude and longitude attributes.	862,781
Arrest	A binary variable that indicates whether a criminal was arrested.	2

Table II shows the attributes selected by the method. The order of occurrence corresponds to the existing correlation between them, according to the approach presented in the previous section. The proposed mean shift method presents as a relevant characteristic a required computation time of 148.95 seconds (see Table II), to select the attributes. This time consists of 17.63 seconds for loading the data set, 107.29 seconds for the creation of clustering, and 24.03 seconds to execute the correlation, allowing the selected attributes to be displayed in minimum processing time.

TABLE II. COMPARISON OF THE ATTRIBUTES SELECTED BY THE METHODS

Method	*Selected Features*	*Time*
Mean Shift	1,10,12,15,17,6,7,8,9	148.95s
RFE	1,4,10,11,12,16,17	2096.49s
Human Expert	1,2,4,5,7,8,9,10	0.00s

In the RFE-based method, 80% of the data set is used for the training of the algorithm, and the remaining 20% of instances of the data set is used for the validation phase required by the method. This method requires 2069.49 seconds to recommend the attributes to be selected (see Table II). The RFE method selects 7 attributes (UICR, Location Description, FBI Code, X coordinate, Y coordinate, Longitude, and Location) from a total of 17.

Additionally, a group of experts was consulted (we call this a human expert-based method), composed of business analysts (employees of the police and criminalistics department) and software engineers who manage public safety systems. The human expert method required 25 hours to analyze the attributes and values of the data set, proposing the following 8 attributes: UICR, Primary Type, Location Description, Domestic, District, Ward, Community Area, and FBI Code.

The mean shift method coincides with the RFE method in 4 proposed attributes to be selected (UICR, FBI Code, Y coordinate, and Location), but only coincide in the position of occurrence of one attribute (UICR), in the list of attributes selected by the methods. On the other hand, the mean shift method coincides with the human expert method in 5 attributes (UICR, District, Ward, Community Area, and FBI Code), and the RFE method agrees with the human expert method only in 3 selected attributes (UICR, Location Description, and FBI Code). The three methods recommend selecting as a first attribute the UICR code, but the order of the rest of the concordant attributes among the methods does not match.

B. J48 Algorithm Results

The J48 decision tree algorithm is used to evaluate and compare the performance of the proposed feature selection algorithm with the RFE method and the human expert method, in terms of predictive accuracy.

The instances of the data subsets used in the experiment were selected and extracted by a random method, automatically, from the original data set. In our experiment a 60-20-20 approach was applied, that is, 60% of the observations were used to train our model, 20% of the instances were used for the test phase of the model and the remaining 20% of records in the data set were used for the validation of the class label prediction model.

We formed the reduced data sets (sub data set) containing those features selected by different feature selection methods applied to the full experimental data set. Then, we trained, tested, and evaluated the J48 classifier on the reduced data sets. The obtained classification accuracies are shown in Table III.

TABLE III. THE ACCURRACY OBTAINED BY J48 ALGORITHM FOR EACH ATTRIBUTE SELECTION METHOD

Algorithm	*Testing Accuracy*	*Evaluation Accuracy*
Mean Shift	0.886173	0.886952
RFE	0.887452	0.887816
Human Expert	0.886638	0.887259

It can be observed that the classifier trained on the mean shift data set tends to exhibit slightly lower classification accuracy in testing phase (0.886173). The classifier trained on the data set containing features selected by the RFE method constantly performed better than the classifier trained with mean shift and human expert data sets, both in the testing phase and evaluation phase.

The next important result that can be observed in Table III is that the mean shift method exhibits an improvement classification performance in the evaluation phase (0.886952), compared to the accuracy achieved in testing phase. Additionally, the mean shift method reduces the distance with the precision obtained by the other two methods.

IV. CONCLUSIONS

Feature selection provides an effective way to solve the dimensionality problem by removing irrelevant and redundant data, which can reduce computation time, improve learning accuracy, and facilitate a better understanding of the learning model or data.

The mean shift method proposed allows obtaining the necessary features without human intervention, because the clustering is carried out automatically without the need to define a K number a priori. The selection of the most representative features is made with the support of Pearson’s linear correlation, pre-defining a threshold of +-0.5, allowing to discard irrelevant attributes.

On the other hand, in the RFE method, it is necessary to define initially how many attributes we want the algorithm to select. In addition, since it is a wrapper-type method, it depends entirely on the learning algorithm with which it was trained.

In our experiment, it is observed that the computation time required by the RFE method is very high compared to the processing time required by the mean shift method. This method only needs a 7.19% of the time of the RFE method to determine the attributes to be selected. Therefore, we consider that in data mining and automatic learning tools, with real-time execution, it is feasible to use the proposed mean shift method, because the computation time required to select the attributes by mean shift method is better than RFE and human expert methods.

Acknowledgment

This work was supported by the National Council of Science and Technology (CONACYT) and Basic Science Research Project SEP-CONACYT of Mexico under Grant 256922.

References

[1] C. Harrison, et al., “Foundations for smarter cities,” IBM Journal of Research and Development, vol. 54, no. 4, July 2010, pp. 1–16.

[2] H. Schaffers, et al., “Smart cities and the future internet: Towards cooperation frameworks for open innovation,” in The Future Internet, J. Domingue, et al., Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 431–446. [Online]. https://doi.org/10.1007/978-3-642-20898-0_31 [retrieved: May, 2019].

[3] S. P. Mohanty, U. Choppali, and E. Kougianos, “Everything you wanted to know about smart cities: The internet of things is the backbone,” IEEE Consumer Electronics Magazine, vol. 5, no. 3, July 2016, pp. 60–70.

[4] C. Yin, et al., “A literature survey on smart cities,” Science China Information Sciences, vol. 58, no. 10, 2015, pp. 1–18.

[5] A. H. Alavi, P. Jiao, W. G. Buttlar, and N. Lajnef, “Internet of things-enabled smart cities: State-of-the-art and future trends,” Measurement, vol. 129, 2018, pp. 589 – 606.

[6] C. M. Sosa-Reyna, E. Tello-Leal, and D. Lara-Alabazares, “Methodology for the model-driven development of service oriented IoT applications,” Journal of Systems Architecture, vol. 90, 2018, pp. 15 – 22.

[7] A. Al-Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari, and M. Ayyash, “Internet of things: A survey on enabling technologies, protocols, and applications,” IEEE Communications Surveys Tutorials, vol. 17, no. 4, 2015, pp. 2347–2376. https://doi.org//10.1109/COMST.2015.2444095 [retrieved: Jun, 2019].

[8] E. Al Nuaimi, H. Al Neyadi, N. Mohamed, and J. Al-Jaroodi, “Applications of big data to smart cities,” Journal of Internet Services and Applications, vol. 6, no. 1, Dec 2015, p. 25. [Online]. https://doi.org/10.1186/s13174-015-0041-5 [retrieved: Jun 2019].

[9] C. Lim, K.-J. Kim, and P. P. Maglio, “Smart cities with big data: Reference models, challenges, and considerations,” Cities, vol. 82, 2018, pp. 86 – 99. [Online]. https://doi.org/10.1016/j.cities.2018.04.011 [retrieved: Apr, 2019].

[10] M. Ge, H. Bangui, and B. Buhnova, “Big data for internet of things: A survey,” Future Generation Computer Systems, vol. 87, 2018, pp. 601 – 614.

[11] C.-W. Tsai, C.-F. Lai, H.-C. Chao, and A. V. Vasilakos, “Big data analytics: a survey,” Journal of Big Data, vol. 2, no. 1, Oct 2015, p. 21. [Online]. https://doi.org/10.1186/s40537-015-0030-3 [retrieved: May, 2019].

[12] A. M. S. Osman, “A novel big data analytics framework for smart cities,” Future Generation Computer Systems, vol. 91, 2019, pp. 620 – 633.

[13] M. Han and W. Ren, “Global mutual information-based feature selection approach using single-objective and multi-objective optimization,” Neurocomputing, vol. 168, 2015, pp. 47 – 54.

[14] A. Jain and D. Zongker, “Feature selection: evaluation, application, and small sample performance,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, no. 2, Feb 1997, pp. 153–158.

[15] J. Cai, J. Luo, S. Wang, and S. Yang, “Feature selection in machine learning: A new perspective,” Neurocomputing, vol. 300, 2018, pp. 70 – 79.

[16] R. Sheikhpour, M. A. Sarram, S. Gharaghani, and M. A. Z. Chahooki, “A survey on semi-supervised feature selection methods,” Pattern Recognition, vol. 64, 2017, pp. 141 – 158.

[17] G. Chandrashekar and F. Sahin, “A survey on feature selection methods,” Computers Electrical Engineering, vol. 40, no. 1, 2014, pp. 16 – 28, 40th-year commemorative issue.

[18] J. Tang, S. Alelyani, and H. Liu, Feature selection for classification: A review. Data Classification: Algorithms and Applications. CRC Press, 2014, pp. 37–64.

[19] M. Moradkhani, A. Amiri, M. Javaherian, and H. Safari, “A hybrid algorithm for feature subset selection in high-dimensional datasets using FICA and IWSSr algorithm,” Applied Soft Computing, vol. 35, 2015, pp. 123 – 135.

[20] S. Solorio-Fernández, J. A. Carrasco-Ochoa, and J. F. Martínez-Trinidad, “A review of unsupervised feature selection methods,” Artificial Intelligence Review, 2019, pp. 1– 42.

[21] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, “Gene selection for cancer classification using support vector machines,” Machine Learning, vol. 46, no. 1, 2002, pp. 389–422. [Online]. https://doi.org/10.1023/A:1012487302797 [retrieved: May, 2019].

[22] D. Comaniciu and P. Meer, “Mean shift: a robust approach toward feature space analysis,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 5, May 2002, pp. 603–619.

[23] M. A. Carreira-Perpinan, Handbook of Cluster Analysis. New York: Chapman and Hall/CRC, 2016, ch. Clustering Methods Based on Kernel Density Estimators: Mean-Shift Algorithms.

[24] A. Tehreem, S. G. Khawaja, A. M. Khan, M. U. Akram, and S. A. Khan, “Multiprocessor architecture for real-time applications using mean shift clustering,” Journal of Real-Time Image Processing, 2017, pp. 1– 17.

[25] Chicago Police Department, “Reported Crime - Public Safety dataset”. [Online]. https://data.cityofchicago.org/Public-Safety/Crimes-2001-topresent/ ijzp-q8t2/data, 2018, [retrieved: Apr, 2019].