New Method for Selecting Exemplars Application to Roadway Experimentation

Emilien Bourdy, Kandaraj Piamrat, Michel Herbin, Hacène Fouchal

New Method for Selecting Exemplars Application to Roadway Experimentation

Emilien Bourdy[1], Kandaraj Piamrat[2(✉)], Michel Herbin[1], and Hacène Fouchal[1]

[1] CReSTIC, Université de Reims Champagne-Ardenne, Reims, France
{emilien.bourdy,michel.herbin,hacene.fouchal}@univ-reims.fr
[2] LS2N, Université de Nantes, Nantes, France
kandaraj.piamrat@univ-nantes.fr

Abstract. Nowadays, data are generated and collected in many domains from various sources. In most of the cases, they are handled as common data where some simple calculations are used to analyse them as measuring the average, the maximum, the deviation, etc. For instance, the average number of children in European families is 1.8 children. This kind of assessment is far away from reality: the number of children should be an integer number. For this reason, exemplars have a finer meaning since its aim, in this case, is to look of an exemplar of a common family in Europe which has 2 children (the most representative family). The aim of this paper is to propose a methodology able to extract representative exemplars from a dataset. This methodology has been experimented with dataset extracted from experimentations of connected vehicle traces. This data analysis has shown some interesting features: the vehicle connectivity guarantees that messages are not lost.

Keywords: Exemplars · Sampling · Data mining
Intelligent Transport Systems

1 Introduction

When we want to study drivers behavior, we have huge amount of data, each corresponds to a behavior. To analyze them, we need to explore these data and we extract exemplars. In order to do this, we propose a new methodology based on an estimation of the local density in a neighborhood of each data. By doing this, we extract representative exemplars that will reflect the different behaviors. We can choose the number of exemplars we want to reduce the processing cost and time. If the classes are known, we try to provide at least one exemplar in each class.

A set of exemplars is a classical way for storing and representing the cognitive structures [1]. The exemplars are real data extracted from a large dataset unlike the prototypes that are artificial data such as the statistics. Thus, the selection of a few exemplars that represent the whole dataset is one of the first step when exploring a dataset. For instance, the selection of exemplars is central

to several clustering methods [2]. The selection of exemplars is a case-oriented process which is also called sampling [3]. The goal is to extract a small subset of representative data from the dataset.

The use of sampling techniques is necessary when the dataset is too large. But sampling techniques are also used when the treatment of each individual data needs for lot of money, time, effort, etc. Moreover the selection of exemplars is essential in all the fields where tests or treatment are impossible to implement on the whole population (i.e. the whole dataset). In such trials it may be risks associated with individual treatment. So ethical reasons involve testing the treatment on only a small sample of the population.

Data are described generally with a large number of variables. Because of the sparsity of high dimensional data space, the selection of exemplars becomes highly difficult when data lies within such a space. The phenomenon is known as the curse of dimensionality [4]. The method we propose uses the approach of parallel coordinates [12] to escape the curse of dimensionality when extracting exemplars.

The area of VANET is a very challenging domain nowadays. It attracts many research teams mainly to prepare the future vehicles, which will probably be connected and autonomous. Connected vehicles exchange a lot of messages and the need of analysis on these large amount becomes very urgent.

In [6], the authors present a formal model of data dissemination within VANETs and study how VANET characteristics, specifically the bidirectional mobility on well defined paths, have an impact on performance of data dissemination. They investigate the data push model in the context of TrafficView, which have been implemented to disseminate information about the vehicles on the road.

In [7], the authors handle two aspects The derivation of real mobility patterns to be used in a VANET simulator and the simulation of VANET data dissemination achieved with different broadcast protocols in real traffic setting.

Most of data analysis are done on one hop sent messages but it could be interesting to analyse data over routing issues as the in [8]. We will first present the sampling method in Sect. 2, the assessment with different random and known examples in Sect. 3, use the methodology with a roadway experimentation in Sect. 4 and a conclusion in Sect. 5.

2 Sampling Method

Let Ω be a dataset with n data defined by

$$\Omega = \{X_1, X_2, X_3, ...X_n\}.$$

The goal of sampling is to select a subset of Ω, which is called the subset Σ where

$$\Sigma = \{Y_1, Y_2, Y_3, ...Y_p\} \text{ with } Y_j \in \Omega.$$

When sampling, p is much smaller than n ($p \ll n$) and Y_j (with $1 \leq j \leq p$) is a representative or exemplar of Ω. This paper describes a new method to select these exemplars.

Our method is based on an estimation of the local density in a neighborhood of each data. The first exemplar we select is the one with the highest local density. Then the nearest neighbors of this exemplar are removed from Ω. We obtain the following exemplars while iterating the process until the dataset is empty.

2.1 Local Density

In this subsection, we explain how we estimate the local density of each data and how we define the nearest neighbors of an exemplar. Finally, we study the number of exemplars we can propose using this sampling method.

In this paper we only consider multidimensional quantitative data. Thus, X_i with $1 \leq i \leq n$ is a vector defined by:

$$X_i = (v_1(i), v_2(i), ...v_p(i))$$

where v_1, v_2,... v_p are the p variables that are the features of data. In this context each data lies a p-dimensional data space.

Classically, the density is defined using a unit hypervolume. For instance, the hypersphere of radius α can define the unit hypervolume. In the data space, the local density at X is then equal to the number of data of Ω lying inside the unit hypersphere centered in X. Unfortunately the definition of density comes up against the curse of dimensionality [4]. When the dimension of the data space increases, the volume of the available data becomes sparse and the classical definition of density has no meaning. For circumventing this drawback, we define the density for each variable using the approach of parallel coordinates [12] (see Fig. 1). Therefore, we have p densities, each defined in a one-dimensional space. The sum of these densities gives us a density-based index that we use in the whole data space.

Let us define the density computed in the one-dimensional space of the variable v_j (with $1 \leq j \leq p$). The dataset Ω is projected in this space and we obtain n values with:

$$\Omega_j = \{v_j(1), v_j(2), v_j(3), ...v_j(n)\}.$$

These values are in the range $[min_j, max_j]$ where $min_j = \min_{1 \leq i \leq n} (v_j(i))$ and $max_j = \max_{1 \leq i \leq n} (v_j(i))$. Let us define the unit interval we use to compute the density at each value x. Let k be an integer between 1 and n. If we expected a local density equal to k, then the length α_j we propose for the unit interval is equal to $\alpha_j = \frac{max_j - min_j}{n} * k$. Thus the local density at x is equal to the number of elements of Ω_j that are in the unit interval $[x - \alpha_j/2,\ x + \alpha_j/2]$. The local density at X_i for the variable v_j is then defined by:

$$density_j(X_i) = \#\{\ [v_j(i) - \alpha_j/2,\ v_j(i) + \alpha_j/2]\ \cap\ \Omega_j\ \}.$$

Finally, the local density at X_i for all the variables is defined by:

$$density(X_i) = \sum_{1 \leq j \leq p} density_j(X_i).$$

We select the data which has the highest local density. This data is the first exemplar of Ω:

$$Y_1 = \arg\max_{X_i \in \Omega} \quad density(X_i).$$

2.2 Nearest Neighbors

The previous procedure enables us to select only one exemplar. We obtain the following exemplars by reducing the dataset and iterating this procedure. The dataset is reduced by removing Y_1 and its nearest neighbors.

Let us describe our definition of the nearest neighbors of a data X in a dataset Ω. The neighbors of X_i for the variable v_j are the data of Ω that are in the unit interval centered in X_i. This neighborhood N_j is defined by:

$$N_j(X_i) = \{X_k \in \Omega \text{ with } v_j(k) \in [v_j(i) - \alpha_j/2,\ v_j(i) + \alpha_j/2]\}.$$

The nearest neighbors of X_i for all the variables should be in the neighborhoods for each variable. Thus the nearest neighbors of X_i are in the neighborhood N defined by:

$$N(X_i) = \bigcap_{1 \leq j \leq p} N_j(X_i).$$

To select the second exemplar Y_2 we exclude the first one Y_1 and its nearest neighbors $N(Y_1)$. We apply the procedure defined in the previous section within a reduced dataset $\Omega \setminus N(Y_1)$. Then Y_2 the data with the highest local density within the reduced dataset.

We iterate the procedure until the reduced dataset is empty. The exemplars we obtain gives us the samples of Ω.

2.3 Number of Exemplars

We set our method of sampling using the parameter k where k is an expected local density at each data. The value of k lies between 1 and n when the dataset has n data. In this section, we explain how the value of k can change the number of exemplars selected through our sampling method.

Let us consider a toy example with 200 simulated data ($n = 200$) with 5 variables ($p = 5$). Figure 1 displays the profiles of these data with 200 dashed broken lines. The exemplars are selected using the parameter value $k = 100$. We obtain 7 exemplars (bold broken lines in Fig. 1).

The number of selected exemplars decreases when the parameter value k increases. Figure 2 shows that the number of selected exemplars decreases from 200 to 1 when the density parameter k increases. This property of our method

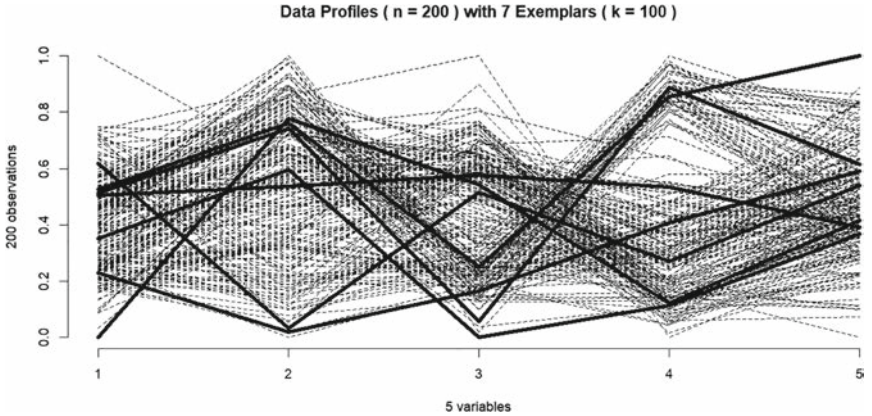

Fig. 1. Profiles of 200 simulated data with 5 variables (dashed lines) election of 7 exemplars with a parameter value k = 100 (bold lines)

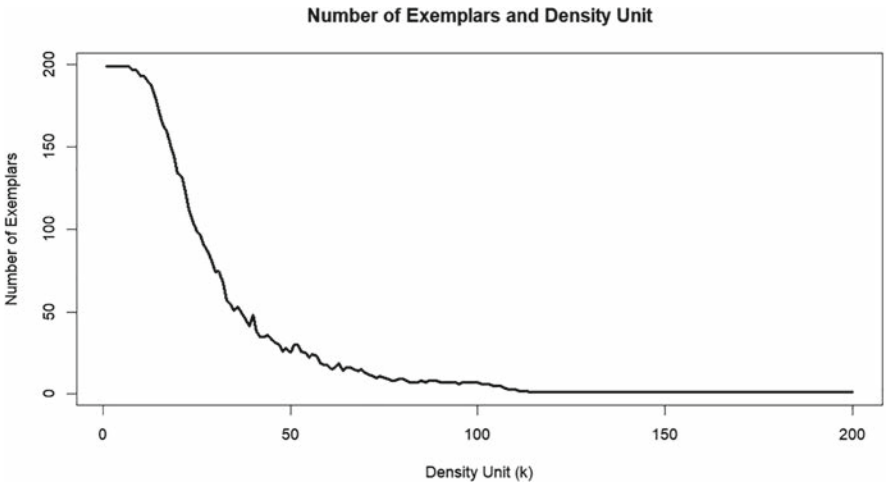

Fig. 2. Number of selected exemplars decreases from 200 to 1 when the density parameter k increases from 1 to 200

enables us to adapt a strategy to select the number of samples that we extract from the dataset. If we want a specific number of samples selected from the initial dataset, then we can adjust the parameter k to obtain the expected number of exemplars.

3 Assessment of Sampling

The exploratory analysis of a dataset is complex for many reasons. The dataset is often divided into classes but the distribution of these classes is unknown.

Moreover, the number of these classes is also unknown. To better understand data, the use of an complementary exploratory trial on a smaller dataset is often necessary. The selection of a reduced number of samples should then represent all the classes of the dataset. For this reason, we will evaluate our sampling method under controlled conditions when the distribution of the classes is known. But of course, the method remains designed for applications in exploratory analysis when the classes are unknown. This method is particularly useful when classes have large overlapping and when the classes have very different numbers of data. In such cases, the classical methods of clustering very often fail.

Let us consider a dataset with known distribution of classes for assessing our sampling method. We verify that the distribution of the selected exemplars between classes remains comparable with the distribution of data within the initial dataset. Table 1 gives the results we obtain with some simulations.

Table 1. Distribution between classes within a dataset ($n = 200$) and within the selected exemplars

Number of classes	Distribution in dataset ($n = 200$)	Number of selected exemplars	Exemplars distribution between classes
4	(42, 51, 65, 42)	25	(4, 8, 8, 5)
4	(42, 51, 65, 42)	18	(5, 6, 4, 3)
4	(42, 51, 65, 42)	13	(3, 3, 5, 2)
4	(42, 51, 65, 42)	9	(2, 3, 3, 1)
4	(42, 51, 65, 42)	7	(2, 2, 2, 1)
4	(7, 103, 62, 28)	10	(1, 3, 4, 2)
5	(40, 47, 55, 9, 49)	10	(2, 3, 2, 2, 1)
6	(6, 12, 76, 80, 24, 2)	9	(2, 1, 2, 1, 2, 1)
7	(23, 16, 51, 46, 1, 36, 27)	8	(1, 1, 1, 2, 0, 1, 2)
8	(37, 9, 3, 19, 48, 12, 45, 27)	9	(3, 0, 0, 1, 1, 1, 3, 1)

In the first five rows of the table, we use the dataset displayed in Fig. 1. This dataset is simulated using four classes with a large overlapping. The 200 data are randomly distributed between these classes. (42, 51, 65, 42) is the distribution between the four classes. The number of selected exemplars decreases when the parameter k increases. We obtain 25, 18, 13, 9 and 7 exemplars using respectively 50, 60, 70, 80 and 100 as values of k. In these five simulations, the four classes are effectively represented by the exemplars. However, when k increases, the number of selected exemplars becomes too small for representing each class.

In the last five rows of Table 1, we simulate five datasets with respectively 4, 5, 6, 7 and 8 classes. The number of data in each class is randomly selected

and it could be very different from one class to another one. The datasets have 200 data and the parameter k is equal to 80 when selecting exemplars. When the number of classes increases, the number of exemplars becomes too small for representing each class (see the two last rows of the table). However, these classes are represented if the number of selected exemplars increases (i.e. if we decrease the value of the parameter k).

Let us study the sampling with real datasets. We consider some datasets of UCI repository (see in [5]). Table 2 displays the selection of exemplars using our blind method (i.e. when the classes are unknown) on the classical dataset called "Iris", "Wine", "Glass", "Haberman" and "Ecoli".

Table 2. Distributions between classes with a real dataset and with selected exemplars ($n =$ number of data, $p =$ number of variables)

Name of dataset	n	p	Distribution in dataset	Number of exemplars	Distribution of exemplars
Iris	150	4	(50, 50, 50)	8	(3, 3, 2)
Wine	178	13	(59, 71, 48)	9	(2, 4, 3)
Glass	214	9	(70, 76, 17, 13, 9, 29)	19	(1, 6, 1, 5, 1, 5)
Haberman	306	3	(225, 81)	10	(7, 3)
Ecoli	336	7	(143, 77, 2, 2, 35, 20, 5, 52)	23	(3,9,0,0,3,3,2,3)

These datasets have respectively 3, 3, 5, 2 and 8 classes. Our sampling method gives generally an exemplar in each classes. Obviously the method fails if the number of classes is high relative to the number of selected exemplars. Moreover, the method often fails if the number of elements within one class is very low. For instance, in the last line of Table 2, two classes have only 2 elements and these classes are not represented by the exemplars. But these classes can be represented by an exemplars if we increase the number of exemplars we select.

4 Roadway Experimentation

In the Scoop@f [9] (Cooperative System at France) project, Intelligent Transport System (ITS) is experimented in the real life. To do that, connected vehicles drive on roadway and communicate with the infrastructure or other vehicles via a specific WiFi called ITS-G5. Messages used in Scoop@f are CAM [10] (Cooperative Awareness Message) and DENM [11] (Decentralized Environmental Notification Message). CAM is an application beacon with information about ITS station position, speed if it's a mobile station, etc. DENM is used to warn about events. In this experimentation, the vehicle drives on a roadway and send DENMs automatically. The event of this experimentation is a slippery road. When vehicle

send DENM, it logs the 30 previous seconds and 30 next seconds. We used theses logs with our methodology.

Theses logs contain 3 201 data of 17 variables (7 for the acceleration control, the steering wheel angle, the strength braking and 8 for the exterior lights). The acceleration control is defined by the brake, gas and emergency brake pedals, collision warning, ACC (Adaptive Cruise Control), cruise control and speed limiter utilization. And the exterior lights are defined by the low and high beam, left and right turn signal (warning is the combination of both), daytime, reverse, fog and parking light.

Here we want to describe characteristics from the experimentation and trying to modeling vehicle behavior on this type of route. By using the 3 201 data we obtain the Fig. 3, we can see that there no big differences. It's explain by the fact that roadway has less changes than urban road. We then used our methodology with the $k \in \{40, 80, 100, 140, 180, 200\}$ that give us the Table 3. With $k = 40$, 142 samples are extracted, 102 with $k = 80$, 92 with $k = 100$, 58 with $k = 140$, 46 with $k = 180$ and 48 with $k = 200$. The reduction of number of samples in comparison with the number of entries is explained by the fact that there is a lot of data that are equals. With our methodology, a big pre-processing is made, dividing by $\{22, 31, 34, 55, 69, 66\}$ the number of data to process.

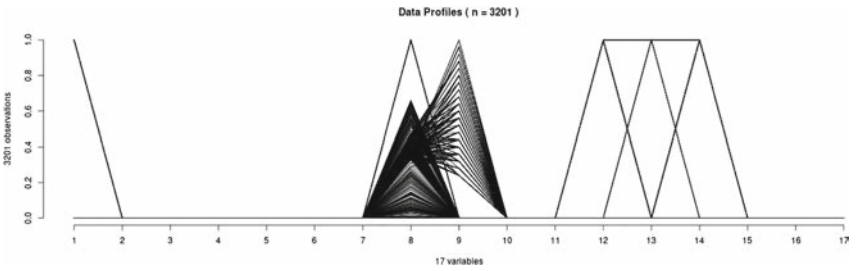

Fig. 3. Profiles of 3 201 data with 17 variables from roadway experimentation.

Table 3. Selection of exemplars with different values of the parameter k from roadway experimentation.

k	Number of exemplars	Division
40	142	22
80	102	31
100	92	34
140	58	55
180	46	69
200	48	66

5 Conclusion and Future Work

In this paper, we have presented a new methodology to select exemplars from a dataset containing multidimensional quantitative data. Our method is based on an estimation of the local density in a neighborhood of each data. With this methodology, it's also possible to select exemplars from classes, and then reduce the number of data in theses classes. This methodology was first used with random data and then with known real data, and with a roadway experimentation from the Scoop@f [9] project to perform exemplars of the situation described by the experimentation.

In the near future, we will use the methodology with other experimentation and developing tools to create classes and representing each experimentation. This work is a contribution for designing tools to analyze data of roadway experimentations in the project Scoop@f.

Acknowledgement. This work was made possible by EC Grant No. INEA/CEF/TRAN/A2014/1042281 from the INEA Agency for the SCOOP project. The statements made herein are solely the responsibility of the authors.

References

1. Frixione, M., Lieto, A.: Prototypes vs exemplars in concept representation. In: International Conference on Knowledge Engineering and Ontology Development, KEOD (2012)
2. Frey, B.J., Dueck, D.: Clustering by passing messages between data points. Science **315**, 972–976 (2007)
3. Cochran, W.G.: Sampling Technique. Wiley Eastern Limited, New Delhi (1985)
4. Houle, M.E., Kriegel, H.-P., Kröger, P., Schubert, E., Zimek, A.: Can shared-neighbor distances defeat the curse of dimensionality? In: Gertz, M., Ludäscher, B. (eds.) SSDBM 2010. LNCS, vol. 6187, pp. 482–500. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-13818-8_34
5. Bache, K., Lichman, M.: UCI Machine learning repository, School of Information and Computer Sciences, University of California, Irvine. http://archive.ics.uci.edu/ml (2013)
6. Nadem T, Shankar P, Iftode, L.: A comparative study of data dissemination models for VANETs. In: 3rd Annual International Conference on Mobile and Ubiquitous Systems (MOBIQUITOUS), July 2006
7. Castellano A., Cuomo F.: Analysis of urban traffic data sets for VANETs simulations. CoRR abs/1304.4350 (2013)
8. Ayaida, M., Barhoumi, M., Fouchal, H. Ghamri-Doudane, Y., Afilal, L.: PHRHLS: a movement-prediction-based joint routing and hierarchical location service for VANETs. In: 2013 IEEE International Conference on Communications (ICC), pp. 1424–1428 (2013)
9. Scoop@f. http://www.scoop.developpement-durable.gouv.fr/
10. CAM: ETSI EN 302 637–2; Intelligent Transport Systems (ITS); Vehicular Communications; Basic Set of Applications; Part 2: Specification of Cooperative Awareness Basic Service. European Standard. ETSI, November 2014

11. DENM: ETSI EN 302 637–3; Intelligent Transport Systems (ITS); Vehicular Communications; Basic Set of Application; Part 3: Specifications of Decentralized Environmental Notification Basic Service. European Standard. ETSI, November 2014
12. Heinrich, J., Weiskopf, D.: State of the art of parallel coordinates. STAR? State of The Art Report, Visualization Research Center, University of Stuttgart, Eurographics (2013)