Supervised pre-processing are useful for supervised clustering

O. Alaoui Ismaili[1,2], V. Lemaire[1], and A. Cornuéjols[2]

[1] Orange Labs, AV. Pierre Marzin 22307 Lannion cedex France
(oumaima.alaouiismaili, vincent.lemaire)@orange.com
[2] AgroParisTech 16, rue Claude Bernard 75005 Paris
antoine.cornuejols@agroparistech.fr

Abstract Over the last years, researchers focus their attention on a new approach that combines the main characteristics of both traditional clustering and supervised classification tasks. This new approach is called by *supervised clustering*. Motivated by the importance of pre-processing approaches in a traditional clustering context, we suppose that a supervised pre-processing step could help traditional clustering to obtain the same supervised clustering tacks. This paper conducts experiments which show that the traditional clustering is competitive comparing to existing supervised clustering algorithms if a supervised pre-processing step is used.

1 Introduction

To organize huge collections of data, clustering algorithms have shown significant results over the past few years. Clustering is an unsupervised learning approach that allows to discover the global structure of data (i.e. clusters). Precisely, given a dataset, it identifies different data subsets which are meaningful (see Figure 1. a)). Clustering outputs are meaningful if clusters are heterogeneous (i.e. inter-similarity) and instances within each cluster share similar features (i.e. intra similarity). This learning problem has motivated a huge body of work and has resulted in a large number of efficient algorithms which only differ in their definition of what is an efficient clustering (Kaufman L. et al. (1990)),(Jain A. et al. (1999)). Clustering has found applicability in numerous real-life application domains such as marketing, city planning, medicine (Berry M. et al.(1997) and Berson A. at al.(1999)) and so forth.

In contrast, classification is a supervised learning approach that is characterized by the presence of additional information named *target class.* The main goal of this approach is to construct a learning model which is able to predict class membership for new instances (see Figure 1. b)). In this setting, several algorithms are developed, e.g, Kotsiantis S. B. gives a review of common classification algorithms in machine learning (Kotsiantis S. B. (2007)).

Recently, researchers focus their attention to combine characteristics of both clustering and classification tasks in the goal to discover the structure of the target class. This advanced research is called *Supervised clustering* (for instance see (Al-Harbi et al. (2006)), (Eick et al. (2004))). The main idea is to construct or modify clustering algorithms at the aim of finding clusters where instances are very likely to belong to the same class. Formally, *Supervised clustering* is a clustering technique where instances in each cluster share both characteristics (homogeneity) and class label. The generated clusters are labeled with the majority class of their instances. Figure 1 illustrates the difference between: clustering, classification and supervised clustering.

Generally, clustering tasks require an unsupervised pre-processing step (for example, see (Milligan, G. et al. (1988)) or for K-means algorithm see (Celebi et al. (2012))), for instance, to prevent features with large ranges from dominating the distance calculations. Motivated by the importance of pre-processing for the traditional clustering, in this paper we attempt to verify the following assumption: does *supervised pre-processing* help traditional clustering in a supervised clustering context? i.e., using a supervised pre-processing step before the traditional clustering could provide an efficient clustering in term of accuracy of predictions?

The remainder of this paper is organized as follows: Section 2 briefly describes some related works about supervised clustering. Section 3 presents some unsupervised preprocessing and two supervised pre-processing approaches. Section 4 verifies the proposed assumption by comparing at first both supervised and unsupervised pre-processing in term of accuracy (ACC) for the resulting clustering. Subsequently, by comparing the traditional clustering using supervised pre-processing step to other supervised clustering algorithms. Finally, a conclusion with some perspectives is presented as a conclusion in the last section.

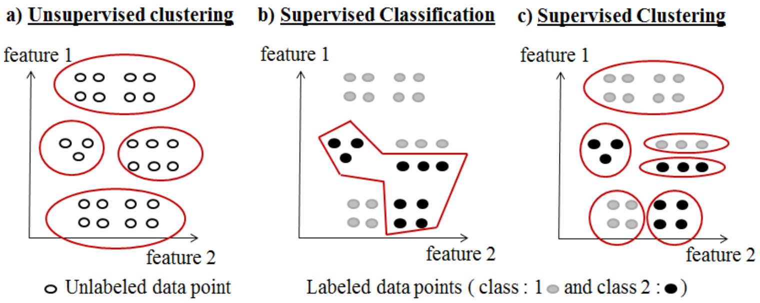
a) Unsupervised clustering
b) Supervised Classification
c) Supervised Clustering
feature 1
feature 2
feature 1
feature 2
feature 1
feature 2
○ Unlabeled data point
Labeled data points (class : 1 ● and class 2 : ●)

Figure 1. Classification processes

2 Related works

In the setting of supervised clustering, many algorithms are developed in a manner to achieve the desired objective: identify heterogeneous clusters where instances within each cluster belong to the same class. In this section some of these algorithms are presented below.

Sinkkonen et al. (2002) proposed in this context a supervised algorithm which use probabilistic approach based on discriminative clustering to minimize distortion within clusters. Finley et al. (2005) proposed an SVM algorithm to train clustering algorithm in the goal to obtain desirable clustering. This algorithm could learn from an item-pair similarity measure to optimize clustering performance with respect to a variety of performance measures. Aguilar et al. (2001) proposed an algorithm called by *S-NN* which employed hierarchical clustering algorithm based on the nearest neighbor technique. Qu et al. (2004) presented supervised model-based clustering algorithm based on multivariate Gaussian mixture model. To estimate different model parameters, the EM algorithm is used. Slonim et al. (1999) and Tishby et al. (1999) presented supervised algorithms based on bottom-up agglomerative approach. Al-Harbi et al. (2006) developed K-means algorithm to be useful in supervised context. The resulting algorithm combines Simulated Annealing with modified K-means algorithm. Eick et al. (2004) introduced three representative-based supervised algorithms. Two of them are greedy algorithms: Single Representative Insertion/Deletion Steepest Decent Hill Climbing with Randomized Start (*SRIDHCR*) and Supervised Partitioning Around Medoids (SPAM). The last one is an evolutionary computing algorithm called Supervised Clustering using Evolutionary Computing (*SCEC*). They also proposed a new fitness function which is used to measure the performance of clustering algorithms. Jirayusakul et al. (2007) introduced two supervised algorithms named Robust Supervised Growing Neural Gas (*RSGNG*) and Supervised Growing Neural Gas (*SGNG*). These algorithms are based on prototype-based clustering methodology. Bungkomkhun at al. (2012) proposed a grid-based supervised clustering algorithm which is based on two crucial methods: grid-based clustering method and bottom-up subspace clustering method.

3 Pre-processing

The following notations are used below:

Let $D = \{(X_i, Y_i)\}_1^N$ denote a training dataset of size N, where $X_i = \{X_{i1}, ..., X_{id}\}$ is a vector of d features and $Y \in \{C_1, ..., C_J\}$ is the target class of size J. Let K denote the number of clusters used in a clustering algorithm.

3.1 Unsupervised pre-processing

Unsupervised pre-processing step is a common requirement for clustering tasks. Several unsupervised pre-processing approaches are developed depending on the nature of features: continuous or categorical.

For continuous features, to the best of our knowledge, data normalization is the most frequently used. It acts to weight the contribution of different features with the aim of making the distance between instances meaningful. Formally, normalization scales each continuous feature in a specific range such that one feature cannot dominate the others. The common data normalization approaches are: *Min-Max*, *statistical* and *rank* normalization.

- *Min-Max Normalization* (NORM) : If the minimum and maximum values are given for each continuous feature, it can be then transformed to fit in the range $[0, 1]$ using the following formula: $X'_{iu} = \frac{X_{iu} - min(X_{iu})}{max(X_{iu}) - min(X_{iu})}$. Where X_{iu} is the original value of feature u . If minimum and maximum values are equal, then X'_{iu} is set to zero.

- *Statistical Normalization* (SN) : This approach transforms data derived from any normal distribution into standard normal distribution with unit variance and null mean. The formula that allows this transformation is : $X'_{iu} = \frac{X_{iu} - \mu}{\sigma}$ where μ is the mean of the feature, σ is its standard deviation.

- *Rank Normalization* (RN): The purpose of rank Normalization is to rank continuous feature values and then scale feature into $[0, 1]$. The different steps of this approach are: $i)$ Rank feature values u from lowest to highest values and then divided it into H intervals, where H is a parameter, often equal to 100. $ii)$Assign for each interval a label $r \in \{1, ..., H\}$. $iii)$ If X_{iu} belong to the interval r, then $X'_{iu} = \frac{r}{H}$.

For categorical features, among the existing approaches of unsupervised pre-processing, we use in this study ***Basical grouping approach (BGB)***. It aims to transform feature modalities into a vector of Boolean values. The different steps of this approaches are: $i)$ grouped feature modalities into g groups with equal frequencies, where g is a parameter, often equal to 10 . $ii)$ Assign for each group a label $r \in \{1, ..., g\}$. $iii)$ A full disjunctive coding is used.

3.2 Supervised pre-processing

In this paper, we suggest that the objective of the supervised representation is to estimate the univariate conditional density ($P(X|C)$). To obtain this estimation a supervised discretization method is used for continuous features and a supervised grouping method is used for categorical ones. There are several methods that could achieve the above objective. In this study, we have used the *MODL* approach. It searches to find the adequate split of the domain in intervals or groups of modalities which give us optimal information about the repartition of data between the different classes. The reader can fin a detailed description about this approach in (Boullé, M.(2006)) and in (Boullé, M.(2005)).

After this supervised transformation, a feature recoding is done to obtain additional information about features distribution. Two types of recoding are presented in this paper. The first one is a supervised recoding called by *conditional info* (C.I). It provides amount information about the feature distribution conditionally to a class label. The second one is an unsupervised recoding called by *Binarization* (BIN).

- Conditional Info: C.I : Each feature m is recoded in a qualitative attribute containing I_m recoding values. So, each instance of data is recoded as a vector of discret modalities $X = X_{1i_1}, X_{2i_2}, \ldots, X_{Mi_M}$. Where X_{mi_m} represents the recoding value of X_m on the feature m with the discrete modality index i_m. After this recoding, all initial attributes are represented as a numeric form. The initial vector containing M attributes (continuous and categorical) becomes a vector containing $M * J$ numeric components: $log(P(X_{mi_m}|C_j))$.

The most remarkable point in this process is that if two instances are close in term of distance, they are close also in term of their class membership. A detailed description of this process exists in (Lemaire V. (2012)).

- Binarization: BIN : In this process, each feature is described on t boolean features. Where t is a number of intervals or groups of modalities generated by MODL or other supervised approach. The synthetic feature takes 1 as a value if the real value of the original feature belong to the corresponding interval or group of modalities and takes zero if not. To measure the similarity between instances, the Hamming distance (Hamming, R. W. (1950)) can be used.

Discussion: Advantages and drawbacks of the proposed approaches

The first step of the proposed approaches aims to give an additional information about the distribution of data, for each feature, in the different class labels. In this study, the MODL approach is used: it is characterized by the its efficiency and robustness. To further enhance the performance of our proposed approaches, we add a second step: recoding.

For Binarization approach, this recoding is based on the full disjunctive coding. It transforms each feature into a vector of boolean features. The size of the vector depends on the number of interval or group of modalities associated with each feature. Hence, the size of the new feature space mainly depends on the number of intervals or groups of modalities for all features. Beside, the similarity between instances is assessed such that similar instances belong to the same interval or group of modalities (e.g. Hamming distance). Nevertheless, this approach does not provide any additional information about the class membership of instances, i.e non supervised recoding.

For Conditional Info, the recoding step provides, for each feature, an amount of information related to the target class. That is by calculating $log(P(X_{mi_m}|C_j))$. This recoding allows to obtain a new feature space of apriori-fixed size which corresponds to the total number of class labels in the dataset. The similarity between instances is interpreted as a Bayesian distance. However, it does not allow to keep the notion of instances: two different instances belonging to different intervals (or groups of modalities) can have equal values of $log(P(X_{mi_m}|C_j))$.

4 Experimentation

In this section, we present and compare at first the average performance of both supervised and unsupervised pre-processing approaches using K-means algorithm. Then, we compare and discuss the average performance of both supervised pre-processing and other supervised clustering algorithms. These experiments are intended to affirm the ability of supervised pre-processing to provide better results than unsupervised pre-processing and also prove the competitiveness of a traditional clustering algorithm (K-means) preceded by a supervised pre-processing step comparing to some supervised algorithms in a supervised clustering context.

4.1 Protocol

To test the validity of our assumption, we choose to use the standard K-means algorithm (MacQueen, J.B.(1967)). To avoid the problem of initialization, for this study, we decide to use the K-means++ algorithm (Arthur, D. et al. (2007)). Therefore, in order to get a *”good performance”* using this initialization technique, we realized 100 different partitions. At this stage, it is important to define what is the best partition? To be consistent with the definition of a *supervised clustering*, we search a criterion that allows us to choose the closest partition to the partition given by the target class. In fact, the main aim is to get a compromise between intra similarity and prediction. The heterogeneity of clusters is granted by the K-means algorithm and the class membership of instances inside each cluster is verified by the chosen criterion. For this, we use ARI[3] (Hubert, L. et al.(1985)) criterion to select the best partition. For pre-processing approaches, we use those presented above in section 3. Table 1 presents a list of these approaches.

Table 1. The used pre-processing approaches

Unsupervised pre-processing		**Supervised pre-processing**	
Cont features	Cat features	Cont features	cat features
RN	BGB	BIN	BIN
CR	BGB	C.I	C.I
NORM	BGB		

To evaluate and compare the behavior of different pre-processing approaches in term of their capacity to help traditional clustering in a supervised context, some tests are performed on different databases from UCI (Blake, C.L. et al (1998)). Table 2 presents the databases used in this study.

In order to compare the obtained results with some supervised clustering algorithms, we do: (i) 20×5 folds for Auto-import, Breast, Contraceptive and Pima datasets (like in (Al-Harbi et al. (2006))). (ii) 10×10 folds for Glass, Heart, Vehicle and Iris datasets (like in (Eick et al. (2004))).

[3] The ARI criterion is calculated between: i) the target class which is considered as a reference and ii) the ID-cluster generated by K-means algorithm

Table 2. The used datasets from UCI

dataset	N	# Var	# Cat	# Cont	dataset	N	# Var	# Cat	# Cont
Auto-import	205	26	11	15	Heart-stat-log	270	13	3	10
Breast cancer	699	9	0	9	Iris	150	4	0	4
Contraceptive	1473	9	7	2	Pima	768	8	0	8
Glass	214	10	0	10	Vehicle	846	18	0	18

4.2 Results

Part 1 : Comparing supervised and unsupervised pre-processing

Table 3 presents the average performance of K-means algorithm in term of predictions (ACC criterion), using each pre-processing approach (section 3.2) for 6 datasets. In this case, the number of clusters is estimated[4]. Based on this value, the ACC is calculated from the corresponding partition in a test dataset. The results in this table show that: (1) supervised pre-processing approaches have most of the time a better performance than unsupervised pre-processing approaches. (2) Binarization (BIN) and Conditional Info (C.I) are close with a small preference for BIN.

In the case where K is given[5] , we obtain also the same result. For example, Figures 2 and 3 represent respectively the case where K is an output and where K is an input for Auto-Import dataset. This result shows clearly the influence of supervised pre-processing steps (blue boxplots) on the K-means performance (using ACC criterion).

Part 2 : Comparing supervised pre-processing to other supervised clustering algorithms

We compare the obtained results using the standard K-means algorithm preceded by a supervised pre-processing step (BIN or C.I) to a supervised K means algorithm proposed by Eick or Al-Harbi. These results are available in (Eick et al. (2004)) and (Al-Harbi et al. (2006)) respectively. Table 4 presents a summary of the average performance of the used methods in term of predictions in the case where K is estimated (Eick) and where K is given (Al-Harbi). The mean results of Eick or Al-Harbi (who performed a single x-fold cross validation) are in the variance of the results using a standard K-means

[4] K is estimated : It is varied from 1 to 64. Then, for each value of K, a x-fold (see section 4.1) cross validation is performed and the mean value of ARI is calculated. Finally, the estimated value is correspond to the closest partition to the partition given by the target class (higher value of ARI versus the value of K in train dataset)

[5] K is given : It equals to a number of class label

Table 3. Average performance of K-means algorithm in term of predictions using several pre-processing approaches. (H= Heart, C = Contraceptive, P = Pima, I= Iris, V= Vehicle and B = Breast

		K	ARI Train	ACC Test			K	ARI Train	ACC Test
H	RN-BGB	2	0.424	0.815 ± 0.067	I	RN-BGB	3	0.676	0.854 ± 0.080
	SN-BGB	2	0.366	0.796 ± 0.076		SN-BGB	3	0.626	0.832 ± 0,089
	NORM-BGB	3	0.299	0.778 ± 0.078		NORM-BGB	3	0.720	0.877 ± 0.079
	BIN-BIN	2	0.461	0.813 ± 0.076		BIN-BIN	4	0.877	0.933 ± 0.064
	C.I-C.I	2	0.461	0.808 ± 0.079		C.I-C.I	3	0.840	0.902 ± 0.083
C	RN-BGB	3	0.076	0.617 ± 0.027	V	RN-BGB	7	0.196	0.546 ± 0.036
	SN-BGB	2	0.055	0.617 ± 0.031		SN-BGB	8	0.157	0.507 ± 0.049
	NORM-BGB	3	0.072	0.612 ± 0.030		NORM-BGB	8	0.159	0.510 ± 0.044
	BIN-BIN	3	0.098	0.632 ± 0.029		BIN-BIN	5	0.256	0.558 ± 0.039
	C.I-C.I	3	0,080	0, 623 ± 0, 032		C.I-C.I	5	0.283	0.589 ± 0.033
P	RN-BGB	2	0.137	0.671 ± 0, 038	B	RN-BGB	2	0.896	0.973 ± 0.012
	SN-BGB	2	0.184	0.706 ± 0.033		SN-BGB	2	0.867	0.965 ± 0.016
	NORM-BGB	5	0.149	0.684 ± 0.035		NORM-BGB	2	0.862	0.964 ± 0.016
	BIN-BIN	2	0.176	0.699 ± 0.043		BIN-BIN	2	0.904	0.974 ± 0.012
	C.I-C.I	2	0.266	0.740 ± 0.033		C.I-C.I	2	0,895	0, 969 ± 0.020

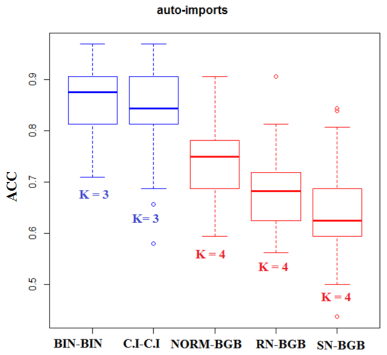

Figure 2. Average performance of K-means (K is an output) using supervised pre-processing (red boxplots) and unsupervised pre-processing (blue boxplots)

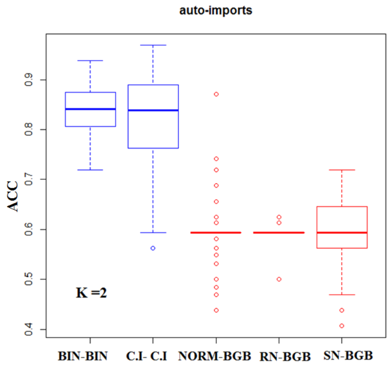

Figure 3. Average performance of K-means (K is an input) using supervised pre-processing (red boxplots) and unsupervised pre-processing (blue boxplots)

preceded by a supervised pre-processing. We can also observed that a standard K-means with a supervised pre-processing step could conserve a lower number of clusters (in Glass dataset, $K = 34, 7$ and 6 for respectively Eick, Binarization and Conditional Info approaches).

Table 4. Comparing with Eick and Al-Harbi algorithms

Comparing with Eick algorithm : (K is an output)						
	Glass dataset		Heart dataset		Iris data set	
	K	ACC Test	K	ACC Test	K	ACC Test
Eick algorithm	34	0.636	2	0.745	3	0.973
K-means with BIN	7	0.677 ± 0.091	2	0.813 ± 0.076	4	0.933 ± 0.064
K-means with C.I	6	0.620 ± 0.093	2	0.808 ± 0.079	3	0.902 ± 0.083

Comparing with AL-Harbi algorithm : (K is an intput)						
	Auto-import dataset		Breast dataset		Pima data set	
	K	ACC Test	K	ACC Test	K	ACC Test
Al-Harbi algorithm	2	0.925	2	0.976	2	0.746
K-means with BIN	2	0.831 ± 0.054	2	0.974 ± 0.012	2	0.699 ± 0.043
K-means with C.I	2	0.814 ± 0.102	2	0.969 ± 0.020	2	0.740 ± 0.033

5 Conclusion

This paper has presented the influence of a supervised pre-processing step on the performance of a traditional clustering (especially K-means) in term of predictions. The experimental results showed the competitiveness of a traditional clustering using a supervised pre-processing step comparing to: i) unsupervised preprocessing approaches. ii) other methods of supervised clustering from the literature (especially Eick and Al-Harbi algorithms). Future works will be done (i) to combine supervised pre-processing presented in this paper with supervised K-means; (ii) to define a better supervised pre-processing approach to combine the advantages of BIN and C.I without their drawbacks.

References

Kaufman L, Rousseeuw P (1990): Finding groups IN DATA: An introduction to cluster analysis.*John Wiley and Sons Inc*

Jain, A. K. and Murty, M. N. and Flynn, P. J. (1999): Data Clustering: A Review. *ACM Comput. Surv. 31(3) (pp.264-323)*

Berry M, Linoff G (1997): Data mining techniques for marketing, sales, and customer support. *John Wiley and Sons, New York*

Berson A, Thearling K, Smith S (1999): Building data mining applications for CRM. *McGraw-Hill Professional Publishing*

Kotsiantis, S. B. (2007): Supervised Machine Learning: A Review of Classification Techniques. *Proceedings of the 2007 Conference on Emerging Artificial Intelligence Applications in Computer Engineering.(pp. 3-24)*

Celebi E. M., Hassan A. Kingravi, Patricio A. Vela (2012): A Comparative Study of Efficient Initialization Methods for the K-Means Clustering Algorithm . *Expert Systems with Applications 40*

Al-Harbi S. H., Rayward-Smith V. J. (2006): Adaptive k-means for supervised clustering, *Applied Intelligence (pp.219-226)*

Eick C.F., Zeidat N., Zhao Z.(2004): Supervised clustering algorithms and benefits. *In proceedings of the 16th IEEE International Conference on Tools with Artificial Intelligence (ICTAI04) , Boca.(pp.774–776)*

Milligan, G. and Cooper, M. (1988) : A study of standardization of variables in cluster analysis *Journal of Classification, Springer-Verlag (pp.181-204)*

Sinkkonen, J., Kaski, S., Nikkil, J. (2002): Discriminative clustering: Optimal contingency tables by learning metrics. *In Elomaa, T., Mannila, H., Toivonen, H., eds.: ECML. Volume 2430 of Lecture Notes in Computer Science., Springer (pp.418–430)*

Finley, T., Joachims, T. (2005): Supervised clustering with support vector machines. *In: Proceedings of the 22Nd International Conference on Machine Learning. ICML '05, New York, NY, USA, ACM (pp.217–224)*

Aguilar-Ruiz, J.S., Ruiz, R., Santos, J.C.R., Girldez, R. (2001): SNN: A supervised clustering algorithm. *In Monostori, L., Vncza, J., Ali, M., eds.: IEA/AIE. Volume 2070 of Lecture Notes in Computer Science., Springer (pp.207–216)*

Qu, Y., Xu, S.(2004): Supervised cluster analysis for microarray data based on multivariate gaussian mixture. *Bioinformatics 20(12)(pp.1905–1913)*

Slonim, N., Tishby, N.a. (1999): Agglomerative information bottleneck. *In Proc. Neural Information Processing Systems (pp.617-623)*

Tishby, N., Pereira, F.C., Bialek, W. (1999): The information bottleneck method. *In: Proc. of the 37-th Annual Allerton Conference on Communication, Control and Computing. (pp.368–377)*

Jirayusakul, A., Auwatanamongkol, S. (2007): A supervised growing neural gas algorithm for cluster analysis. *Int. J. Hybrid Intell. Syst. 4(2) (pp.129–141)*

Bungkomkhun, P., Auwatanamongkol, S. (2012): Grid-based supervised clustering algorithm using greedy and gradient descent methods to build clusters. *IJCSI International Journal of Computer Science Issues, 9*

Boullé, M.(2006): MODL: a Bayes optimal discretization method for continuous attributes. *Journal of Machine Learning 65(1) (pp.131–165)*

Boullé, M.(2005): A Bayes optimal approach for partitioning the values of categorical attributes. *Journal of Machine Learning Research,(pp.1431-1452)*

Lemaire, V., Clérot, F., Creff, N. (2012):K-means clustering on a classifier-induced representation space : application to customer contact personalization. *Annals of Information Systems, Springer, Special Issue on Real-World Data Mining Applications.*

Hamming, R. W. (1950) : Error detecting and error correcting codes. *Bell system technical journal (pp. 147-160)*

MacQueen, J.B.(1967): Some methods for classification and analysis of multivariate observations. *In Cam, L.M.L., Neyman, J., eds.: Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability. University of California Press (pp.281–297)*

Arthur, D., Vassilvitskii, S. (2007) : K-means++: The advantages of careful seeding. *In: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms. SODA '07 (pp.1027–1035)*

Hubert, L., Arabie, P. (1985): Comparing partitions. *Journal of classification 2(1) (pp.193–218)*

Blake, C.L., Merz, C.J. (1998): Uci repository of machine learning databases. http://archive.ics.uci.edu/ml/.