# Online Cluster Approximation via Inequality

Shriprakash Sinha

Technische Universiteit Delft, Pattern Recognition and Bioinformatics Lab,
EWI Bldg., Mekelweg 4, 2628 CD Delft, The Netherlands
Shriprakash.Sinha@gmail.com

**Abstract.** Given an example-feature set, representing the information context present in a dataset, is it possible to reconstruct the information context in the form of clusters to a certain degree of compromise, if the examples are processed randomly without repetition in a sequential online manner? A general transductive inductive learning strategy which uses constraint based multivariate Chebyshev inequality is proposed. Theoretical convergence in the reconstruction error to a finite value with increasing number of (a) processed examples and (b) generated clusters, respectively, is shown. Upper bounds for these error rates are also proved. Nonparametric estimates of these error from a sample of random sequences of example set, empirically point to a stable number of clusters.

## 1 Introduction

*This work focuses on approximating the number of clusters in an online unsupervised learning framework with no prior knowledge of the number of clusters.* The current work achieves this using a transductive-inductive (TI) learning strategy. Motivation for the proposed work comes from the recent works on conformal learning theory (CLT) [1] and [2], where TI has been widely used to estimate the prediction of an unlabeled example based on the already processed data. *Also, the estimation of clusters using a multidimensional data is an open area of research.* Finding motivation from [3] on $k$-means clustering using the Euclidian ball multivariate Chebyshev inequality (MCI) and [4] on image representation via a hybrid model based on space filling curve, the proposed work exploits a generalization of MCI [5], to capture the interactions among examples with multiple features. MCI provides bounds for multidimensional data which is afflicted by the curse of dimensionality that make estimation of multivariate probabilities a difficult task. A generalization of MCI is the consideration of probability content of a multivariate normal random vector to lie in an Euclidean $n$-dimensional ball [6]. This work employs a conservative approach of using the Euclidian $n$-dimensional ellipsoid to restrict the spread of the probability content [7]. In abstract terms, let $X$ be a stochastic variable in $\mathcal{N}$ dimensions with a mean $E[X]$. Further, let $\Sigma$ be the covariance matrix of all examples, each containing $\mathcal{N}$ features and $\mathcal{C}_p \in \mathcal{R}$ (MCI parameter), then the MCI in [7] states that: $\mathcal{P}\{(X - E[X])^T \Sigma^{-1} (X - E[X]) \geq \mathcal{C}_p\} \leq \frac{\mathcal{N}}{\mathcal{C}_p}$ and can be transformed into

L. Iliadis et al. (Eds.): AIAI 2012 Workshops, IFIP AICT 382, pp. 176–181, 2012.

$\mathcal{P}\{(X - E[X])^T \Sigma^{-1}(X - E[X]) < \mathcal{C}_p\} \geq 1 - \frac{\mathcal{N}}{\mathcal{C}_p}$ i.e. the probability of the spread of the value of $X$ around the sample mean $E[X]$ being greater than $\mathcal{C}_p$, is less than $\mathcal{N}/\mathcal{C}_p$.

*The novelty of the current work is in combining the TI learning strategy with the MCI, as the algorithm processes the examples in an online unsupervised paradigm.* Before presenting the details, a succinct view of why and how the task is accomplished is given: • The algorithm aims to reconstruct the information context by building clusters, as it processes a random sequence of unlabeled examples. The number of clusters is not known apriori. • A compromise level is used to decide how much loss of information can be tolerated by imposing constraint on the homogeneity of the generated clusters. The compromise level is captured by $\mathcal{C}_p$. $\mathcal{C}_p$ shares similar but not the same concept of defining confidence level as $\varepsilon$ does in the CLT paradigm. The level of confidence generated in the proposed algorithm is $1 - \frac{\mathcal{N}}{\mathcal{C}_p}$ as opposed to $1 - \varepsilon$ in CLT. • The quality of the prediction in CLT is checked based on the p-values generated online. The proposed algorithm generates reconstruction error online as a measure of the quality of reconstruction. The reconstruction error is computed as follows: (1) All processed examples present in a cluster are assigned mean feature values by averaging examples across each feature. (2) For each example, Euclidian distance between the newly assigned mean feature values and the original feature values is computed. (3) Summing the individual deviations and averaging over all examples gives the total reconstruction error. The computed reconstruction error changes dynamically as new examples are processed. • Dependence on a sequence makes the algorithm as weak learner. To resolve this, probability distribution of reconstruction errors generated from a sample of random sequences is estimated. Empirically, it is found that with a maximum probability value, there exists a low reconstruction error value to which the algorithm converges. This low reconstruction error points to the adequate number of clusters also. Lastly, the algorithm currently does not work on the merging of the clusters. Section 2 discusses the algorithm in detail followed by implications and analysis of convergence in section 3. Section 4 discusses empirical results followed by conclusion.

## 2 Transductive-Inductive Learning Algorithm

Let the examples $(z_i = x_i)$ be sampled from the example set randomly without repetition in a sequential manner, thus forming a sequence $\mathcal{Z}_i$. The algorithm works in alternative steps by (a) creating new clusters using 1-Nearest Neighbour (NN) transductive learning and (b) learning the association of a new example to an existing cluster by evaluating the MCI (i.e inductive learning). In case no associations are found to exist, a new cluster is created by employing 1-Nearest Neighbour (NN) transductive learning. Thus the algorithm starts with no clusters at all and uses NN to initialize a new cluster. Once a cluster is initialized (say with $x_i$ and $x_j$), the size of the cluster depends on the number of examples getting associated with it. The MCI controls the degree of uniformity of

different feature values of examples that constitute the cluster. The association of the example to a cluster happens as follows: Let the new random example (say $x_t$) be considered for checking the association to a cluster. If the spread of example $x_t$ from $\mathbf{E}_q(x)$ (the mean of the $q^{th}$ cluster $\{x_i, x_j\}$ and $x$ is set of all examples in $q^{th}$ cluster), factored by the covariance matrix $\Sigma_q$, is below $\mathcal{C}_p$, then $x_t$ is considered as a part of the cluster. Using MCI, it boils down to: $\mathcal{P}\{(x_t - \mathbf{E}_q[x_i, x_j])^T \Sigma_q^{-1} (x_t - \mathbf{E}_q[x_i, x_j]) \geq \mathcal{C}_p\} \leq \frac{\mathcal{N}}{\mathcal{C}_p}$ or $\mathcal{P}\{(x_t - \mathbf{E}_q[x_i, x_j])^T \Sigma_q^{-1} (x_t - \mathbf{E}_q[x_i, x_j]) < \mathcal{C}_p\} \geq 1 - \frac{\mathcal{N}}{\mathcal{C}_p}$ Satisfaction of this criterion suggests a possible cluster to which $x_t$ could be associated. This test is conducted for all the existing clusters. If there are more than one cluster to which $x_t$ can be associated, then the cluster which shows the minimum deviation from the new random point is chosen. Once the cluster is chosen, its size is extended to one more example i.e. $x_t$. The cluster now constitutes $\{x_i, x_j, x_t\}$ and its $\mathbf{E}_q(x)$ and $\Sigma_q$ recomputed. In case of failure to find any association, the algorithm employs 1-NN transductive algorithm to find a closest neighbour (in unseen data), of the current example under consideration. This neighbour together with the current example forms a new cluster. Thus the step of cluster formation or association is repeated online to reconstruct the information content in the dataset.

## 3 Implications

*In this paper, the term decomposition is synonymous to cluster.* The proposed work uses MCI and the probability associated with it to define a decomposition as follows:

**Definition 1.** *Let the $q^{th}$ cluster $\mathcal{D}_q$ be a* ***decomposition****, then:* $\mathcal{D}_q = \{x_i | \forall i \; ((x_i - \mathbf{E}_q(x))^T \; \Sigma_q^{-1} \; (x_i - \mathbf{E}_q(x))) < \mathcal{C}_p\}$

A decomposition expands by testing a new point $x_t$ via the inequality $(x_t - \mathbf{E}_q(x))^T \; \Sigma_q^{-1} \; (x_t - \mathbf{E}_q(x)) < \mathcal{C}_p$. The probability of the satisfaction of this inequality is given by $\mathcal{P}\{(x_t - \mathbf{E}_q(x))^T \; \Sigma_q^{-1} \; (x_t - \mathbf{E}_q(x)) < \mathcal{C}_p\} \geq 1 - \frac{\mathcal{N}}{\mathcal{C}_p}$. Thus at any point in time, while processing the dataset online, the probability of existence of $\mathcal{D}_q$ is lower bounded by a value of $1 - (\mathcal{N}/\mathcal{C}_p)$.

**Lemma 1.** *For any cluster $q$, the probability of existence of $\mathcal{D}_q$ is lower bounded by a value of $1 - (\mathcal{N}/\mathcal{C}_p)$.*

It is important to note that $\mathcal{C}_p$ plays a major role in deciding whether a new example $x_t$ belongs to $\mathcal{D}_q$. If the new example cannot be associated to a particular decomposition, then it is tested with other decompositions using MCI with same value of $\mathcal{C}_p$. Thus $\mathcal{C}_p$ acts as a constraint in checking the homogeneity of $\mathcal{D}_q$. Now, since all examples (except for those used to from new clusters where the NN is used) are tested using MCI and finally associated with one or the other decomposition for which the constraint is satisfied, the total number of generated clusters is limited. This limitation is enforced indirectly via the $\mathcal{C}_p$. In this scenario, it can be expected that the total number of clusters is upper bounded by $\mathcal{M}/\mathcal{C}_p$ ($\mathcal{M}$ being the total number of examples in the dataset).

**Lemma 2.** *The value of $\mathcal{C}_p$ reduces the initial representation of information content present in $\mathcal{M}$ examples to a representation of information content present in the decompositions whose number is upper bounded by $\mathcal{M}/\mathcal{C}_p$.*

Since the existence of a homogeneous decomposition is bounded probabilistically, the reconstruction error associated with all the examples present in the decomposition are also bounded. For all decompositions, the summation of reconstruction errors is also bounded. Thus for a particular value of $\mathcal{C}_p$ a proof of convergence is needed for the error rates as the number of processed examples and the number of clusters increase.

Two error rates are computed as the random sequence of examples get processed. Let $x_i \in \mathcal{R}^{\mathcal{N}}$ ($\mathcal{N}$ is the number of features) be in the dataset. Since the example is assigned to a particular decomposition $\mathcal{D}_q$, it gets a value of the mean of the all examples that constitute the decomposition. Thus the reconstruction error for the example turns out to be $||x_i - \mathbf{E}_q(x)||_2$. For each cluster $q$, the reconstruction error is $Err_{\mathcal{D}_q} = \sum_{i=1}^{n} ||x_i - \mathbf{E}_q(x)||_2$ ($n$ is the number of examples in the $q^{th}$ cluster). As new examples are processed based on the conceptualized information from the previous examples, the total error computed at after processing the first $pt_{cntr}$ examples in a random sequence is $Err_{val} = \sum_{q=1}^{cluster_{cntr}} Err_{\mathcal{D}_q}$ ($cluster_{cntr}$ is the total number of clusters generated after the $pt_{cntr}$ examples have been processed). The error rate for these $pt_{cntr}$ examples is $Err_1 = Err_{val}/pt_{cntr}$. Finally, the rate of error after every new cluster formation is also computed. This error is denoted by $Err_2$ (i.e. $Err_2 = Err_{val}/cluster_{cntr}$).

**Theorem 1.** *From a dataset, if examples are selected randomly and processed online in a sequential manner without repetition for a particular value of $\mathcal{C}_p$ using the TI, then the reconstruction error rate $Err_1$ converges asymptotically with a probabilistically lower bound or confidence level of $1 - \mathcal{N}/\mathcal{C}_p$ or greater.*

*Proof.* Since $\mathcal{C}_p$ defines level of compromise in information content via lemma 2 and the decompositions $\mathcal{D}_q$ is almost homogeneous, all examples that constitute a decomposition have similar feature values. Due to this similarity between the feature values, the non-diagonal elements of the covariance matrix in the inequality above approach to zero or smaller values. Thus, $\Sigma_q^{-1}$ approaches a diagonal matrix. Multiplying a vector with diagonal matrix scales the vector by some constant factor (say $\mathcal{S}_q$). Thus, if $\Sigma_q^{-1} \approx \mathcal{S}_q \times \mathbf{I}$, were $\mathbf{I}$ is the identity matrix then the inequality equates to: $(x_t - \mathbf{E}_q(x))^T \mathcal{S}_q \mathbf{I}(x_t - \mathbf{E}_q(x)) \lessapprox \mathcal{C}_p \Rightarrow (x_t - \mathbf{E}_q(x))^T \mathbf{I}(x_t - \mathbf{E}_q(x)) \lessapprox \frac{\mathcal{C}_p}{\mathcal{S}_q} \Rightarrow ||x_t - \mathbf{E}_q(x))||_2 \lessapprox \frac{\mathcal{C}_p}{\mathcal{S}_q}$

Thus, if $x_i = x_t$ was the last example to be associated to a decomposition, the reconstruction error $||x_i - \mathbf{E}_q(x)||_2$ for that example would be upper bounded be $\frac{\mathcal{C}_p}{\mathcal{S}_q}$. Consequently, the total error after processing $pt_{cntr}$ examples is also upper bounded, i.e. $Err_{val} = \sum_{q=1}^{cluster_{cntr}} Err_{\mathcal{D}_q} = \sum_{q=1}^{cluster_{cntr}} \sum_{i=1}^{n} ||x_i - \mathbf{E}_q(x)||_2 \lessapprox \sum_{q=1}^{cluster_{cntr}} \sum_{i=1}^{n} \frac{\mathcal{C}_p}{\mathcal{S}_q}$. Thus the error rate $Err_1 = Err_{val}/pt_{cntr}$ is also upper bounded. Different decompositions may have different $\Sigma_q^{-1}$, but in the worst

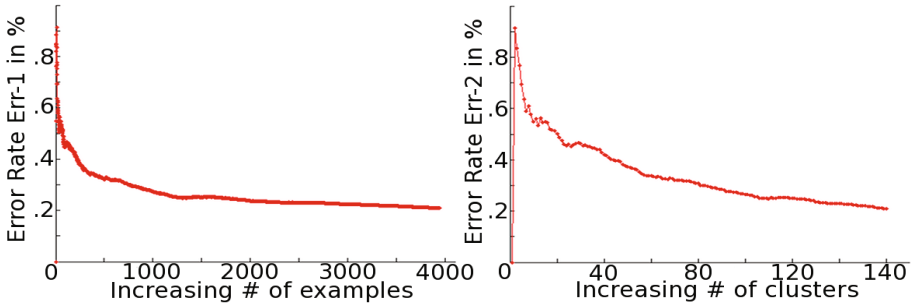


**Fig. 1.** Error rate on left(right) - $Err_1(Err_2)$ for a particular sequence representing a $64 \times 64$ patch of starfish image [4] with increasing number of processed examples(clusters) with $\mathcal{C}_p = 7$

case scenario, if the decomposition with the smallest constant factor is substituted for every other decompositions, then the upper bound on the error is $\frac{\mathcal{C}_p \times \Sigma_{q=1}^{cluster_{cntr}} \Sigma_{i=1}^{n}(1)}{\mathcal{S}_{smallest} \times pt_{cntr}}$ which equates to $\frac{\mathcal{C}_p}{\mathcal{S}_{smallest}}$. Since $\forall q$, $\mathcal{D}_q$s are probabilistically bounded by the Chebyshev inequality, the error rate converges with the probability associated with the Chebyshev inequality. Q.E.D

## 4 Empirical Results

The converging error rate $Err_1$ is depicted in left of figure 1 for a $64 \times 64$ patch in [4] of starfish image. The error rate $Err_2$ is the computation of error after each new cluster is formed. The proof for upper bound on $Err_2$ follows on similar lines. For $\mathcal{C}_p = 7$, the error rate $Err_2$ is depicted on right in figure 1 for same common patch. Intuitively, it can be seen that both the reconstruction error rates converge to an approximately similar value.

Hitherto, reconstruction error and the number of clusters is dependent on a sequence presented to the learner. This points to the problem of whether an image can be reconstructed at a particular $\mathcal{C}_p$ where there is a high probability of finding a low reconstruction error and the number of clusters, from a sample of sequences. The existence of such a probability value would require the knowledge of the probability distribution of the reconstruction error over increasing (1) number of examples and (2) number of clusters generated. KDE is used to estimate the probability distribution of the reconstruction error $Err_1$ and $Err_2$. The KDE is based on a normal kernel using a window parameter that is a function of the number of points. The density is evaluated at 100 equally spaced points that cover the range of the data. The KDE empirically point to the least error rates with high probability. It was found that the error rates $Err_1$, $Err_2$ and the number of clusters, all converge to a particular value (33.1762, 35.9339 and 38, respectively), for a given image (figure 2).

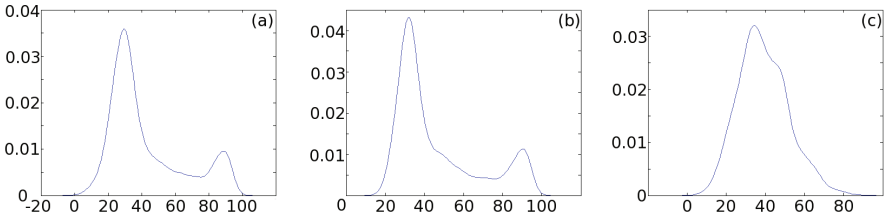


**Fig. 2.** The probability density estimates for (a) $Err_1$ (b) $Err_2$ and (c) the number of clusters generated over 1000 random sequences representing a $64 \times 64$ patch of starfish image [4] with $\mathcal{C}_p = 10$

## 5 Conclusion

Given a random sequence of examples which are processed in an online sequential manner, it is possible to converge on a reconstruction of the information content of the whole dataset to a certain degree compromise with low reconstruction error using the proposed algorithm.

**Acknowledgement.** The author thanks Dmitry Adamskiy Phd candidate and Dr. Ilia Nouretdinov research assistant at the CLRC, Royal Holloway University of London and Dr. David M. J. Tax at the PRB Lab, Delft University of Technology, The Netherlands and the anonymous reviewers for their valuable and critical feedbacks on improving the manuscript.

## References

1. Vovk, V., Gammerman, A., Shafer, G.: Algorithmic learning in a random world. Springer (2005)
2. Shafer, G., Vovk, V.: A tutorial on conformal prediction. The Journal of Machine Learning Research 9, 371–421 (2008)
3. Kanungo, T., Mount, D.M., Netanyahu, N.S., Piatko, C.D., Silverman, R., Wu, A.Y.: An efficient-means clustering algorithm: Analysis and implementation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 881–892 (2002)
4. Sinha, S., Horst, G.: Bounded multivariate surfaces on monovariate internal functions. In: IEEE International Conference on Image Processing, pp. 1037–1040. IEEE Signal Processing Society (2011)
5. Berge, P.O.: A note on a form of tchebycheff's theorem for two variables. Biometrika 29(3-4), 405 (1938)
6. Monhor, D., Takemoto, S.: Understanding the concept of outlier and its relevance to the assessment of data quality: Probabilistic background theory. Earth, Planets, and Space 57(11), 1009–1018 (2005)
7. Chen, X.: A new generalization of chebyshev inequality for random vectors. arXiv:math.ST, 0707(0805v1), 1–5 (2007)