Hybrid K-Means: Combining Regression-Wise and Centroid-Based Criteria for QSAR

Robert Stanforth[1,2], Evgueni Kolossov[1] and Boris Mirkin[2]

[1] ID Business Solutions
2 Occam Court, Guildford, GU2 7QB, UK, {*RStanforth, EKolossov*}*@id-bs.com*

[2] School of Computer Science, Birkbeck, University of London
London WC1E 7HX, UK, *mirkin@dcs.bbk.ac.uk*

Abstract. This paper further extends the 'kernel'-based approach to clustering proposed by E. Diday in early 70s. According to this approach, a cluster's centroid can be represented by parameters of any analytical model, such as linear regression equation, built over the cluster. We address the problem of producing regression-wise clusters to be separated in the input variable space by building a hybrid clustering criterion that combines the regression-wise clustering criterion with the conventional centroid-based one.

1 Introduction

This paper addresses the issues emerging in regression-wise prediction when the sample is not homogeneous or the dependence between the response and input variables is not linear. This type of problem emerges, for example, in the quantitative analysis of relationships between structural features of chemical compounds and their biological activities; this field of research is conventionally referred to as Quantitative Structure-Activity Relationships (QSAR). In such a situation traditional methods of cluster-analysis such as K-Means clustering may not work very well because they capture overall similarities rather than those related to the prediction. In early 70s, E. Diday proposed that a similar way of carrying out cluster analysis can be performed in such situations too (see, for example, Diday (1974)). The nature of the 'centroid' must just be redefined in such a way that any analytical data model, including those of regression or principal component analyses, can become the 'centroid', or 'kernel' of a cluster of entities under consideration (Diday (1974, 1989)).

It is exactly this approach that we are going to pursue for building a clustering better suited for prediction. There is an issue in using the regression-wise clustering for predicting compound activities: the clusters are built in the augmented space of input-response variables, but prediction is to be made based on only the input variables. When, in a typical situation, projections of the clusters to the input variables space overlap, the determination of which of the regression models to apply to an observation may become of an issue. To address this, we further advance in the kernel-based approach to combine the regression-wise with conventional centroid-based clustering so that the clusters found may be more separated in the space of input variables. The combined clustering criterion is referred to as the hybrid K-means criterion here. To assure that no overlaps may occur at all, we supplement the hybrid model with a post-processing option involving one iteration of the centroid-based K-Means applied to the results of the hybrid model in the input variables space so that the resulting clusters are indeed separated in the input space. Another post-processing option involves application of the conventional K-means until convergence.

We present experimental results showing that such a modification indeed reduces the prediction error and find that there is an intermediate value of the hybrid model mixing coefficient leading to the best results.

The remainder is organised as follows. The hybrid criterion is introduced in section 2, after the conventional and regression-wise K-means clustering are defined. Our extension of K-Means methods to the hybrid model is described in section 3. Section 4 presents experimental results and conclusions based on them.

2 The Hybrid K-Means Criterion

We first present a brief recap on the formulation of the K-means algorithm, in preparation for deriving the variants that we shall use. The K-means family of algorithms iteratively optimise a model (of the dataset under consideration) as K clusters. This cluster model comprises a membership element, assigning each member of the dataset to one of the clusters, and a centroid element, which describes each cluster. Iteration of the algorithm proceeds by alternately optimising memberships (leaving centroids fixed) and optimising centroids (leaving memberships fixed).

The optimisation within the K-means algorithm is performed according to a loss function or 'criterion' to be minimised. In the standard 'distance-wise' formulation of K-means in the linear space of some finite feature set V, the loss function is the summary squared Euclidean distance from each point $\mathbf{x}_i$ in the dataset to the centroid $\mathbf{c}_{k(i)}$ of its assigned cluster $k(i)$.

$$L_{dist}(X, C, k) = \sum_{i=1}^{N} \sum_{v \in V} (x_{i,v} - c_{k(i),v})^2 \tag{1}$$

At each step the minimisation of this loss function can be solved directly. With the membership function k fixed, the optimal $c_{k,v}$ is the mean value of $x_{i,v}$ over those points for which $k(i) = k$; in other words the optimal $\mathbf{c}_k$ is the centroid of cluster k, justifying the terminology. On the other hand, with the centroids fixed, the optimal cluster $k(i)$ to which point $\mathbf{x}_i$ may be assigned is the cluster k whose centroid $\mathbf{c}_k$ is closest to $\mathbf{x}_i$. The algorithm terminates when the loss function fails to decrease, so the cluster model has 'converged' to a local (although not in general global) optimum. (Termination will necessarily occur eventually because there are only finitely many configurations of the membership function k.)

Variants of the K-means algorithm can be constructed by introducing different loss functions (Diday (1974)). To remain within the 'alternating optimisation' spirit of the K-means algorithm, we consider loss functions of the following form:

$$L(X, C, k) = \sum_{i=1}^{N} l(\mathbf{x}_i, \mathbf{c}_{k(i)}) \tag{2}$$

This formulation encompasses standard distance-wise K-means via taking l to be the squared Euclidean distance. Note that in general, however, the generalised 'centroids' $\mathbf{c}_k$ need not lie in the same space as the data points $\mathbf{x}_i$ as explained in Diday (1974, 1989).

Regression-wise K-means fits perfectly within this form. This variant of K-means applies to data points $\mathbf{x}_i$ in the linear space of some feature set V as before, but augmented with an associated output or 'activity' value y_i. The intention is that the activity values will be modelled as functions of the feature values x_v. Instead of approximating each point in a cluster k by the cluster's centroid $\mathbf{c}_k$, we model the cluster using a linear regression model $y \approx \sum_v a_{k,v} x_v + b_k$. We then use a squared-error loss function, measuring the summary squared distance along the activity component in augmented feature-activity space from each point to its cluster's regression hyperplane:

$$L_{reg}([X, \mathbf{y}], [A, \mathbf{b}], k) = \sum_{i=1}^{N} (y_i - (\mathbf{a}_{k(i)}^T \mathbf{x}_i + b_{k(i)}))^2 \tag{3}$$

With the membership function k fixed, the optimal cluster regression models $[\mathbf{a}_k, b_k]$ can again be computed directly, in this case by solving the following linear system (which is none other than the normal equations for multivariate linear least squares regression; see, for example, Tabachnik and Fidell (2006)):

$$\begin{aligned} \sum_{i:k(i)=k} \mathbf{x}_i \mathbf{x}_i^T \mathbf{a}_k \; + \sum_{i:k(i)=k} \mathbf{x}_i b_k \; &= \sum_{i:k(i)=k} \mathbf{x}_i y_i \\ \sum_{i:k(i)=k} \mathbf{x}_i^T \mathbf{a}_k \; + \sum_{i:k(i)=k} b_k \; &= \sum_{i:k(i)=k} y_i \end{aligned} \tag{4}$$

As usual, with the generalised 'centroids' fixed, the optimal membership assignment k is that which assigns each point $[\mathbf{x}_i, y_i]$ to the cluster k minimising the loss $(y_i - (\mathbf{a}_{k(i)}^T \mathbf{x}_i + b_{k(i)}))^2$, i.e. the cluster whose regression hyperplane is closest (along the activity axis).

If (for a cluster k) the linear system (4) turns out to be singular, for example because the size of the cluster has fallen below the number of features, then the only option is to 'dissolve' the cluster: its generalised centroid $[\mathbf{a}_k, b_k]$ is left undefined, and it is excluded from the pool when cluster memberships are reassigned in the next and subsequent iterations.

It is straightforward to see that K-means criteria are additive in the sense that, given two loss functions of the above form, their sum is also a valid K-means criterion of this form, distributing over the contributions from each point in the dataset. We may then define 'hybrid K-means' to be K-means clustering performed according to the following combined loss function:

$$L_{hyb} = (1 - p)L_{dist} + pL_{reg} \tag{5}$$

3 Methods

Regression-wise K-means can be viewed as training a composite model for activity (y) values in terms of the feature values ($\mathbf{x}$). The model is composite in the sense that, on each cluster, a separate linear model is computed to be applied on that cluster.

This approach is particularly useful if the activity depends on the feature values via a number of distinct mechanisms, with different mechanisms applying in different regions of feature space. It can also be useful if activity depends on the feature values in a non-linear fashion: the regression-wise clustering will effectively partition the model's non-linear hypersurface in augmented feature-activity space into approximately linear regions.

It should therefore, in principle, be possible for such a composite model resulting from regression-wise K-means to be used for prediction of activity for new points in feature space (whose activity value is not a priori known). Applying the composite model to a new point $\mathbf{x}$ would consist of the following steps:

1. **Classification:** Determine the cluster k to which $\mathbf{x}$ should belong.
2. **Evaluation:** Evaluate the predicted activity as $y = \mathbf{a}_k^T \mathbf{x} + b_k$ according to the regression model for cluster k.

The difficulty with this approach, when based on regression-wise clustering, lies in the classification step. Determination of cluster membership according to the regression-wise K-means criterion (3) is defined for a point $[\mathbf{x}, y]$ in the augmented space, but this dependence on y is circular as y is the unknown we are trying to predict in the first place.

The essence of the problem is that the clusters are defined in augmented space, and so can overlap substantially when projected onto feature space. The solution is to use the hybrid K-means criterion defined in the previous section: the algorithm will then run with a dominant element of distance-wise K-means, promoting separation of clusters in feature space, but retaining a contribution (proportion p) of the regression-wise criterion to guide the clusters towards regions of linearity.

To enable the prediction-time classification in feature space only, we can then follow the hybrid K-means (once it has converged) with one additional iteration with $p = 0$, i.e. according to the distance-wise criterion only. This supplementary iteration – updating memberships then updating centroids/models – will guarantee that the cluster partitioning is defined in terms of feature space only (with no dependence on activity), and that the cluster-specific linear regressions are optimal for the clusters thus defined.

An alternative resolution to compare would be to follow the hybrid K-means (again once it has converged) with as many additional iterations with $p = 0$ as are required until it converges again. This is effectively pure distance-wise K-means, but with hybrid K-means run as a preprocessing step; this is an attempt to orient the initial clusters towards regions of feature space on which separate linear models for activity apply.

Regression-wise and hybrid K-means share with standard distance-wise K-means the requirement for an initial cluster assignment k.

We propose that this initialisation be achieved using Anomalous Pattern Clustering (which also determines the number K of clusters to use), as incorporated into the so-called Intelligent K-Means Algorithm by Mirkin (2005). This Anomalous Pattern Clustering, which itself makes use of a variant of 2-Means to extract the initial clusters, should be applied using the standard distance-only criterion L_{dist}.

4 Results

Ten datasets, each with 5000 points in ten-dimensional feature space augmented with one activity component, were generated randomly. Each dataset was generated with an underlying structure of five clusters, with the clusters' sizes chosen uniformly at random within the simplex of possible relative sizes. Each cluster was assigned a randomly generated mean and spread tensor, based on which the cluster contents were generated according to the multivariate normal distribution. Each cluster was also randomly assigned a linear activity model and an activity error variance; activity values for the points in the cluster were generated according to this linear model with random perturbations according to the error variance.

Each dataset was clustered according to the hybrid K-means algorithm using the criterion derived in section 2, the clustering having first been initialised according to Anomalous Pattern Clustering. Results were output at this stage, and again after one supplementary iteration of K-means, the Minimum distance assignment, was performed with no regression-wise contribution. The K-means algorithm was then allowed to proceed with no regression-wise contribution until convergence was achieved again, after which the results were output for a third and final time.

This procedure was repeated (for each dataset) for several values of p, the relative proportion of the regression-wise contribution.

At each stage, the following results were generated:

1. Regression-wise criterion, expressed as an explained proportion:
$$1 - L_{reg}/L_{reg(worst)}$$
2. Hybrid criterion, expressed as an explained proportion:
$$1 - L_{hyb}/L_{hyb(worst)}$$
3. Distance-wise criterion, expressed as an explained proportion:
$$1 - L_{dist}/L_{dist(worst)}$$
4. Mean relative error of prediction: mean value of
$$|y_{i;predicted} - y_i|/y_i$$
over all structural features i, where the predicted value is according to the regression model of the structure's cluster in the current configuration.

In the above, the ‘worst’ configuration (used for normalising the criterion values) is that obtained using a single cluster and a constant (flat) regression model, leading to the maximum (worst) possible value of the criterion.

Table 1 below presents the mean relative errors of prediction for all datasets at all three stages, for the various values of p under consideration. Mean values over all ten datasets are also included.

The prediction results for the 'original' hybrid K-means (top value in each cell) show a strong decreasing trend (i.e. improvement) as p starts to increase from zero. This is unsurprising, as the relative prediction error closely corresponds to the regression-wise K-means criterion (3). Note that this stage's ‘prediction’ results have a somewhat artificial advantage as they are based on a cluster assignment that in turn depends on prior knowledge of the activity values. Even so, as p continues to increase towards 50%, the decreasing trend in prediction errors is not maintained (and even reversed for several datasets), suggesting that a relentlessly large regression-wise contribution is not aiding the modelling, and that retaining a distance-wise contribution is significantly beneficial in divining the underlying structure of the dataset.

As we would expect, performing the supplementary distance-only iteration of K-means causes the predictive results (centre value in each cell) to worsen. This is because we are now effectively forgoing our 'unfair' prior knowledge of the activity values and basing the cluster selection on feature values and cluster centroids alone. Here we observe, for most of the datasets (and for the mean), a trend in which the predictive power improves as p starts to increase from zero then worsens again as p becomes too large. For any dataset, a value of p specific to that dataset should then be chosen to minimise the prediction errors, expressing the optimal trade-off between regression-wise guidance and distance-wise cluster separation.

	Mean Relative Prediction Error					
Dataset	p=0	p=0.1	p=0.2	p=0.3	p=0.4	p=0.5
1	2.0865	1.5636	1.3256	0.6849	0.5302	0.5673
	2.0865	2.0125	1.9633	1.9708	2.0743	2.1780
	2.0865	2.0839	2.0534	2.0534	2.0560	2.0560
2	0.8609	0.5667	0.5353	0.5582	0.5423	0.5381
	0.8609	0.9819	0.8698	1.0281	0.9231	0.9405
	0.8609	0.8909	0.8902	0.7512	0.7509	0.9148
3	0.4919	0.4072	0.4075	0.4104	0.4080	0.3762
	0.4919	0.5639	0.5516	0.5514	0.5533	0.5389
	0.4919	0.4919	0.4919	0.4919	0.4919	0.4919
4	0.5529	0.4333	0.4219	0.4185	0.4197	0.4200
	0.5529	0.5528	0.5628	0.5153	0.5456	0.5564
	0.5529	0.5528	0.5628	0.5327	0.5330	0.5328
5	0.3150	0.1817	0.1864	0.1747	0.1545	0.1539
	0.3150	0.3199	0.3152	0.3202	0.3149	0.3157
	0.3150	0.3200	0.3200	0.3201	0.3201	0.3200
6	0.8154	0.5500	0.3196	0.3512	0.3738	0.3651
	0.8154	0.8012	0.5979	0.5677	0.5954	0.5942
	0.8154	0.8214	0.5770	0.4339	0.4048	0.5913
7	0.7504	0.4859	0.3332	0.4152	0.5957	0.5252
	0.7504	0.5733	0.5748	0.5082	0.5879	0.6339
	0.7504	0.5743	0.5539	0.5583	0.5814	0.5814
8	0.3790	0.2697	0.2257	0.2110	0.2088	0.2026
	0.3790	0.4102	0.4276	0.4605	0.4666	0.4322
	0.3790	0.3964	0.3955	0.3997	0.3960	0.3584
9	0.1625	0.1607	0.1624	0.1628	0.1633	0.2831
	0.1625	0.1625	0.1625	0.1625	0.1605	0.2176
	0.1625	0.1625	0.1625	0.1625	0.1625	0.1625
10	1.5186	0.5604	0.5875	0.4545	0.4115	0.5007
	1.5186	0.9123	0.9341	0.8754	0.9154	1.0462
	1.5186	0.8755	0.9003	0.9253	0.9155	0.9140
Mean	0.7933	0.5179	0.4505	0.3841	0.3808	0.3932
	0.7933	0.7290	0.6959	0.6960	0.7137	0.7453
	0.7933	0.7170	0.6908	0.6629	0.6612	0.6923

Table 1. Mean Relative Prediction Errors at different values p. Three reals in each cell present: original error of the hybrid model (top), that after one distance-wise iteration (middle), and the error of the hybrid model post-processed with the distance-wise K-means until convergence (bottom).

The alternative scheme of carrying through the supplementary distance-only K-means until convergence is achieved again yields similar, even slightly better, results. (See bottom value in each cell.) The point at which continuing to increase the proportion p of regression-wise contribution starts to have a detrimental effect tends to occur later than it did with only a single supplementary distance-only iteration (around 0.4 rather than 0.3). This can be explained by the fact that performing a greater amount of distance-based post-processing is better able to overcome a heavier regression-wise bias in the initial processing.

Overall, the following conclusions can be made from these experiments:

1. The proposed hybrid-based method indeed allows for a significant, 10%-20%, reduction of the relative prediction error. On average, the error decreases from 79% at only the centroid-based K-Means to 66%.
2. On average, the option of post-processing with the conventional centroid-based K-Means works better. However, when the error of the hybrid model is high (as at datasets 1 and 10), the option of applying the Minimum distance rule once only leads to better results.
3. The best reduction of the error is achieved with the value of the compromise coefficient p at about 0.3.

Table 2 presents the values of the regression-wise, hybrid, and distance-wise K-means criteria (averaged over the ten datasets) at the three stages of analysis. The values in this table demonstrate the degree to which the distance-wise criterion is boosted, to the detriment of the regression-wise criterion, as the supplementary distance-only K-means iterations are performed.

	Values of the criteria at					
Criterion	p=0	p=0.1	p=0.2	p=0.3	p=0.4	p=0.5
regression-wise	0.9777	0.9954	0.9962	0.9965	0.9967	0.9967
	0.9777	0.9806	0.9798	0.9792	0.9783	0.9783
	0.9777	0.9812	0.9828	0.9816	0.9817	0.9825
hybrid	0.6533	0.9672	0.9824	0.9880	0.9910	0.9927
	0.6533	0.9548	0.9676	0.9719	0.9736	0.9751
	0.6533	0.9555	0.9707	0.9744	0.9771	0.9794
distance-wise	0.6533	0.6388	0.6239	0.6108	0.5979	0.5814
	0.6533	0.6512	0.6494	0.6461	0.6455	0.6442
	0.6533	0.6531	0.6529	0.6514	0.6540	0.6540
Mean predictive error	0.7933	0.5179	0.4505	0.3841	0.3808	0.3932
	0.7933	0.7290	0.6959	0.6960	0.7137	0.7453
	0.7933	0.7170	0.6908	0.6629	0.6612	0.6923

Table 2. Values of the criterion of each of the considered models, centroid-wise, regression-wise and the hybrid one, at different values p. Three reals in each cell present: the criterion value after runnig the hybrid K-Means to convergence then stopping (top), that after a supplementary step of one distance-wise iteration (middle), and the error of the hybrid model post-processed with the distance-wise K-means until convergence (bottom).

Obviously our conclusions are based on a rather limited set of experiments. In the future, we are going to, first, extend the simulated data models to other common distributions and, second, apply the hybrid model to real data.

References

DIDAY, E. (1974): Optimization in non-hierarchical clustering. *Pattern Recognition 6 (1), 17-33.*

DIDAY, E., CELEUX, G., GOVAERT, G., LECHEVALLIER, Y., and RALAMBONDRAINY, H. (1989): *Classification Automatique des Données.* Dunod, Paris.

MIRKIN, B. (2005): *Clustering for Data Mining: A Data Recovery Approach.* Chapman & Hall/CRC, Boca Raton, Fl.

TABACHNICK, B.G. and FIDELL, L.S. (2006): *Using Multivariate Statistics (5th Edition).* Allyn & Bacon.