A Latent Variable Pairwise Classification Model of a Clustering Ensemble

Vladimir Berikov

Sobolev Institute of mathematics, Novosibirsk State University, Russia
berikov@math.nsc.ru
http://www.math.nsc.ru

Abstract. This paper addresses some theoretical properties of clustering ensembles. We consider the problem of cluster analysis from pattern recognition point of view. A latent variable pairwise classification model is proposed for studying the efficiency (in terms of "error probability") of the ensemble. The notions of stability, homogeneity and correlation between ensemble elements are introduced. An upper bound for misclassification probability is obtained. Numerical experiment confirms potential usefulness of the suggested ensemble characteristics.

Keywords: clustering ensemble, latent variable model, misclassification probability, error bound, ensemble's homogeneity and correlation.

1 Introduction

Collective decision-making based on a combination of simple algorithms is actively used in modern pattern recognition and machine learning. In last decade, there is growing interest in clustering ensemble algorithms [1,2]. In the ensemble design process, the results obtained by different algorithms, or by one algorithm with various parameters settings are used. After construction of partial clustering solutions, a final collective decision is built.

Modern literature on clustering ensembles can be roughly divided into several main categories. There are a great deal of works in which the ensemble methodology is adapted to new application areas such as magnetic resonance imaging, satellite images analysis, analysis of genetic sequences etc (see, for example, [3,4,5]). Another direction aims to develop clustering ensembles methods of general usage and elaborate efficient algorithms using various optimization techniques (e.g., [6]). Other categories of works are of more theoretical nature; their purpose is to study the properties of clustering ensembles, improve measures of ensemble quality, suggest the ways to achieve the best quality (e.g., [7,8,9]).

There is a large number of experimental evidences confirming a significant raise in stability of clustering decisions for ensemble algorithms (see, for example, [2,10]). At the same time, theoretical grounds of clustering ensembles algorithms, as opposed to the pattern classifier ensembles theory (e.g., [11]), are still in

C. Sansone, J. Kittler, and F. Roli (Eds.): MCS 2011, LNCS 6713, pp. 279–288, 2011.

the early development stage. Existing works consider mainly the asymptotic properties of clustering ensembles (e.g., [7]).

Cluster analysis problems are characterized by the complexity of formalization caused by substantially subjective nature of grouping process. For the definition of clustering quality it is necessary to apply additional a priori information in terms of "natural" classification, to use data generation models etc. In the given work we attempt to corroborate clustering ensemble methodology by utilizing of a pattern recognition model with latent class labels. To avoid problems with class renumbering, a pairwise classification approach is used.

The rest of the paper is organized as follows. In the next section we give basic definitions and introduce the model of ensemble cluster analysis. In the third section we receive an upper bound for error probability (in classifying a pair of arbitrary objects according to latent variable labels) and give some qualitative consequences of the result. In the forth section we introduce the estimates of ensemble characteristics. The next section describes numerical experiment that demonstrates the usage of these notions. The conclusion summaries the work and gives possible future directions.

2 Ensemble Model

Let us consider a sample $s = \{o^{(1)}, \ldots, o^{(N)}\}$ of objects independently and randomly selected from a general population. The purpose of the analysis is to group the objects into $K \geq 2$ classes in accordance with some clustering criterion; the number of classes may be either given beforehand or not (in the latter case an optimal number of classes should be determined automatically).

Let each of the objects be characterized by variables $X_1, \ldots, X_n$. Denote by $x = (x_1, \ldots, x_n)$ the vector of these variables for an object o, $x_j = X_j(o)$, $j = 1, \ldots, n$.

In many clustering tasks it is allowable to consider that there exists a ground truth (latent, directly unobserved) variable

$$Y \in \{1, \ldots, K\}$$

that determines to which class an object belongs. Suppose that the observations of k-th class are distributed according to the conditional dencity function $p_k(x) = p(x|Y = k)$, $k = 1, \ldots, K$.

Consider the following model of data generation. Let each object be assigned to class k in accordance with a priori probabilities $P_k = \mathbb{P}(Y = k)$, $k = 1, \ldots, K$, where $\sum_{k=1}^{K} P_k = 1$. After the assignment, an observable value of x is determined with use of $p_k(x)$. This procedure is repeated independently for each object.

For an arbitrary pair of different objects $a, b \in s$, their correspondent observations are denoted by $x(a)$ and $x(b)$. Let

$$Z = \mathbb{I}(Y(a) \neq Y(b)),$$

where $\mathbb{I}(\cdot)$ is indicator function. Denote by $P_Z = \mathbb{P}[Z = 1|x(a), x(b)]$ the probability of the event "a and b belong to different classes, given $x(a)$ and $x(b)$":

$$P_Z = 1 - \mathbb{P}[Y(a) = 1|x(a)]\,\mathbb{P}[Y(b) = 1|x(b)] - \ldots$$
$$- \mathbb{P}[Y(a) = K|x(a)]\,\mathbb{P}[Y(b) = K|x(b)] = 1 - \sum_{k=1}^{K} \frac{p_k(x(a))p_k(x(b))P_k^2}{p(x(a))p(x(b))},$$

where $p(x(o)) = \sum\limits_{k=1}^{K} p_k(x(o))P_k$, $o = a, b$.

Let a clustering algorithm μ be run to partition s into K subsets (clusters). Because the numberings of clusters do not matter, it is convenient to consider the equivalence relation, i.e. to indicate whether the algorithm μ assigns each pair of objects to the same class or to different classes. Let

$$h_\mu(a, b) = \mathbb{I}[\mu(a) \neq \mu(b)].$$

Let us consider the following model of *ensemble clustering*. Suppose that algorithm μ is randomized, i.e. it depends from a random value Ω from a given set of allowable values (parameters or more generally "learning settings" such as bootstrap samples, order of input objects etc). In addition to Ω, the algorithm's decisions are dependent from the true status of the pair a, b (i.e., from Z):

$$h_\mu(a, b) = h_{\mu(\Omega)}(Z, a, b).$$

Hereinafter we will denote $h_{\mu(\Omega)}(Z, a, b) = h(\Omega, Z)$.

Suppose that

$$\mathbb{P}[h(\Omega, Z) = 1|Z = 1] = \mathbb{P}[h(\Omega, Z) = 0|Z = 0] = q,$$

i.e. the conditional probabilities of correct decision (either partition or union of objects a,b) coincide. One can say that q reflects the *stability* of algorithm under various learning settings. We shall suppose that $q > 1/2$; it means that algorithm μ provides better clustering quality than just random guessing. In machine learning theory, such a condition is known as the condition of *weak learnability*.

Denote $P_h = \mathbb{P}[h(\Omega, Z) = 1]$. This quantity shows the *homogeneity* of algorithm's decisions: P_h close to 0 or 1; or *homogeneity index*

$$I_h = 1 - P_h(1 - P_h)$$

close to 1 means high agreement between the solutions. Note that

$$P_h = \mathbb{P}[h(\Omega, Z) = 1|Z = 1]P_Z + \mathbb{P}[h(\Omega, Z) = 1|Z = 0](1 - P_Z) =$$
$$qP_Z + (1 - q)(1 - P_Z).$$

Suppose that algorithm μ is running L times under randomly and independently chosen settings. As a result, we get random decisions $h(\Omega_1, Z)$, ...,

$h(\Omega_L, Z)$. By $\Omega_1, \ldots, \Omega_L$ we denote independent statistical copies of a random vector Ω.

For every Ω_l, algorithm μ is running independently (it does not use the results obtained with other $\Omega_{l'}$, $l' \neq l$). Suppose that the decisions are conditionally independent:

$$\mathbb{P}[h(\Omega_{i_1}, Z) = h_{i_1}, \ldots, h(\Omega_{i_j}, Z) = h_{i_j} | Z = z] = \\ \mathbb{P}[(h(\Omega_{i_1}, Z) = h_{i_1} | Z = z] \cdot \cdots \cdot \mathbb{P}[h(\Omega_{i_j,Z}) = h_{i_j} | Z = z],$$

where $\Omega_{i_1}, \ldots, \Omega_{i_j}$ are arbitrary learning settings, $h_{i_1}, \ldots, h_{i_j}, z \in \{0, 1\}$ (we shall assume that L is odd).

Let $P_{h,h} = \mathbb{P}[h(\Omega', Z) = 1, h(\Omega'', Z) = 1]$, where Ω', Ω'' have the same distribution as Ω, and $\Omega' \neq \Omega''$. It follows from the assumptions of independence and stability that

$$P_{h,h} = \mathbb{P}[h(\Omega', Z) = 1, h(\Omega'') = 1 | Z = 1] P_Z + \\ \mathbb{P}[h(\Omega', Z) = 1, h(\Omega'') = 1 | Z = 0](1 - P_Z) = \\ q^2 P_Z + (1 - q)^2 (1 - P_Z). \quad (1)$$

Denote $\bar{H} = \frac{1}{L} \sum_{l=1}^{L} h(\Omega_l, Z)$. The function

$$c(h(\Omega_1, Z), \ldots, h(\Omega_L, Z)) = \mathbb{I}[\bar{H} > \frac{1}{2}]$$

shall be called *the ensemble solution* for a and b, based on the majority voting.

For constructing a final ensemble clustering decision, various approaches can be utilized [2]. For example, it is possible to use a methodology based on the pairwise dissimilarity matrix $\mathbb{H} = \left(\bar{H}(o^{(i_1)}, o^{(i_2)}\right)$, where $o^{(i_1)}, o^{(i_2)} \in s$, $o^{(i_1)} \neq o^{(i_2)}$. This matrix can be considered as a matrix of pairwise distances between objects and used as input information for a dendrogram construction algorithm to form a sample partition on a desired number of clusters.

3 An Upper Bound for Misclassification Probability

Let us consider the *margin [11]* of the ensemble:

$$mg = \frac{1}{L}\{ \text{ number of votes for } Z - \text{number of votes against } Z\},$$

where $Z = 0, 1$. It is easy to show that the margin equals:

$$mg = mg(\bar{H}, Z) = (2Z - 1)(2\bar{H} - 1).$$

Using the notion of margin, one can represent the probability of wrong prediction of the true value of Z:

$$P_{err} = \mathbb{P}_{\Omega_1, \ldots, \Omega_L, Z}[mg(\bar{H}, Z) < 0].$$

It follows from the Tchebychev's inequality that

$$\mathbb{P}[U < 0] < \frac{\mathrm{Var}U}{(\mathrm{E}U)^2},$$

where $\mathrm{E}U$ is population mean, $\mathrm{Var}U$ is variance of random value U (it is required that $\mathrm{E}U > 0$). Thus,

$$\mathbb{P}_{\Omega_1,\ldots,\Omega_L,Z}[mg(\bar{H}, Z) < 0] < \frac{\mathrm{Var}\, mg(\bar{H}, Z)}{(\mathrm{E}\, mg(\bar{H}, Z))^2},$$

provided that $\mathrm{E}\, mg(\bar{H}, Z) > 0$.

Theorem. The expected value and variance of the margin are:

$$\mathrm{E}\, mg(\bar{H}, Z) = 2q - 1,$$

$$\mathrm{Var}\, mg(\bar{H}, Z) = \frac{4}{L}\left(P_h - P_{h,h}\right).$$

Proof. We have:

$$\mathrm{E}\, mg(\bar{H}, Z) = \mathrm{E}\,(2Z-1)(\frac{2}{L}\sum_l h(\Omega_l, Z) - 1) = \\ \frac{4}{L}\sum_l \mathrm{E}\, Zh(\Omega_l, Z) - 2\mathrm{E}Z - \frac{2}{L}\sum_l \mathrm{E}\, h(\Omega_l, Z) + 1.$$

Because all $h(\Omega_l, Z)$ are distributed in the same way as $h(\Omega, Z)$, we get:

$$\mathrm{E}\, mg(\bar{H}, Z) = 4\mathrm{E}\, Zh(\Omega, Z) - 2P_Z - 2\mathrm{E}\, h(\Omega, Z) + 1 = \\ 4\mathbb{P}[Z = 1, h(\Omega, Z) = 1] - 2P_Z - 2\mathbb{P}[h(\Omega, Z) = 1] + 1.$$

$$\text{As } \mathbb{P}[h(\Omega, Z) = 1] = \mathbb{P}[Z = 1, h(\Omega, Z) = 1] + \mathbb{P}[Z = 0, h(\Omega, Z) = 1] = \\ qP_Z + (1-q)(1-P_Z) = 2qP_Z + 1 - q - P_Z,$$

we obtain:

$$\mathrm{E}\, mg(\bar{H}, Z) = 4qP_Z - 2P_Z - 2(2qP_Z + 1 - q - P_Z) + 1 = 2q - 1.$$

Consider the variance of margin:

$$\mathrm{Var}\, mg(\bar{H}, Z) = \mathrm{Var}\,(4Z\bar{H} - 2\bar{H} - 2Z) = \mathrm{E}\,(4Z\bar{H} - 2\bar{H} - 2Z)^2 - \\ (\mathrm{E}\,(4Z\bar{H} - 2\bar{H} - 2Z))^2 = \mathrm{E}\,(16Z^2\bar{H}^2 + 4\bar{h}^2 + 4Z^2 - 16Z\bar{H}^2 - 16Z^2\bar{H} + 8Z\bar{H}) - \\ (\mathrm{E}\, mg(\bar{H}, Z) - 1)^2 = \mathrm{E}\,(4\bar{H}^2 + 4Z - 8Z\bar{H}) - 4(1-q)^2$$

(we apply $Z^2 = Z$). Next, we have

$$\mathrm{E}\,\bar{H}^2 = \frac{1}{L^2}\mathrm{E}\left(\sum_l h(\Omega_l, Z)\right)^2 = \frac{1}{L^2}\mathrm{E}\left(\sum_l h^2(\Omega_l, Z)\right) + \frac{1}{L^2}\sum_{\substack{l',l'':\\ l'\neq l''}} \mathrm{E}\left(h(\Omega_{l'}, Z)h(\Omega_{l''}, Z)\right) =$$

$$\frac{1}{L}\mathrm{E}\,h(\Omega, Z) + \frac{L-1}{L}\sum_{\substack{\Omega',\Omega'':\\ \Omega'\neq\Omega''}} \mathrm{E}\left(h(\Omega', Z)h(\Omega'', Z) = \frac{P_h}{L} + \frac{L-1}{L}P_{h,h}.\right.$$

From this, we obtain:

$$\mathrm{Var}\, mg(\bar{H}, Z) \;=\; 4\frac{P_h}{L} \;+\; 4\frac{L-1}{L}P_{h,h} \;+\; 4P_Z \;-\; 8qP_Z \;-\; 4(1\;-\;q)^2.$$

Using (1) finally we get:

$$\mathrm{Var}\, mg(\bar{H}, Z) \;=\; 4\frac{P_h}{L} \;+\; 4\frac{L-1}{L}P_{h,h} \;-\; 4P_{h,h} \;=\; \frac{4}{L}\left(P_h - P_{h,h}\right).$$

The theorem is proved.

Evidently, the requirement $\mathrm{E}\, mg(\bar{H}, Z) > 0$ is fulfilled if $q > 1/2$.

Let us consider the correlation coefficient ρ between $h' = h(\Omega', Z)$ and $h'' = h(\Omega'', Z)$, where $\Omega' \neq \Omega''$. We have

$$\rho = \rho_{h',h''} = \frac{P_{h,h} - P_h^2}{P_h(1 - P_h)}.$$

Because $P_h - P_{h,h} = P_h - P_h^2 + P_h^2 - P_{h,h}$, we obtain

$$\mathrm{Var}\left(mg(\bar{H}, Z)\right) = \frac{4}{L}\left(1 - \rho\right)P_h(1 - P_h).$$

Note that $P_h - P_{h,h} = q(1 - q)$, and after necessary transformations we get the following upper bound for error probability:

$$P_{err} < \frac{1}{L}\left(\frac{1}{1 - 4(1 - \rho)P_h(1 - P_h)} - 1\right).$$

The obtained expression allows to make some qualitative conclusions. Namely, if the model assumptions are fulfilled and $q > 1/2$, then under other conditions being equal the following statements are valid:

- the probability of error decreases with an increase in number of ensemble elements;
- an increase in homogeneity of the ensemble and raise of correlation between its outputs reduce the probability of error (note that a signed value of the correlation coefficient is meant).

4 Estimating Characteristics of a Clustering Ensemble

To evaluate the quality of a clustering ensemble, it is necessary to estimate ensemble's characteristics (in our model – homogeneity and correlation) from a finite number of ensemble elements. For an arbitrary pair of different objects a and b, the estimate of the ensemble's homogeneity can be found as follows:

$$\hat{I}_h(a,b) = 1 - \hat{P}_h(a,b)(1 - \hat{P}_h(a,b)),$$

where

$$\hat{P}_h(a,b) = \frac{1}{L}\sum_{l=1}^{L} h_l(a,b).$$

Unfortunately, the straightforward estimation of the correlation coefficient $\rho(a,b)$ is impossible: under a fixed sample, every pair of clustering algorithms give conditionally independent decisions. Let us introduce a similar notion: the averaged correlation coefficient, where the averaging is done over all pairs of different objects:

$$\bar{\rho} = \frac{cov(h', h'')}{\sigma^2(h)};$$

where the covariance

$$cov(h', h'') = \overline{h'h''} - \bar{h}^2,$$

$$\overline{h'h''} = \frac{2}{N(N-1)}\frac{2}{L(L-1)}\sum_{\substack{a,b:\\ a\neq b}}\sum_{\substack{l',l'':\\ l'\neq' l''}} h_{l'}(a,b)\, h_{l''}(a,b),$$

$$\bar{h} = \frac{2}{N(N-1)}\frac{1}{L}\sum_{\substack{a,b:\\ a\neq b}}\sum_{l} h_l(a,b) = \frac{2}{N(N-1)}\sum_{\substack{a,b:\\ a\neq b}} \hat{P}_h(a,b),$$

and the variance

$$\sigma^2(h) = \overline{h^2} - \bar{h}^2 = \bar{h} - \bar{h}^2.$$

Similarly, it is possible to introduce the averaged homogeneity index:

$$\bar{I}_h = \frac{2}{N(N-1)}\sum_{\substack{a,b:\\ a\neq b}} \hat{I}_h(a,b).$$

5 Numerical Experiment

To verify the applicability of the suggested methodology for the analysis of clustering ensemble behavior, the statistical modeling approach was used. In the modeling, artificial data sets are repeatedly generated according to certain distribution class (a type of "clustering tasks"). Each data set is classified by the ensemble algorithm. The correct classification rate, averaged over a given number of trials, determines algorithm's performance for the given type of tasks.

The following experiment was performed. In each trial, two classes are independently sampled according to the Gauss multivariate distributions

$$\mathcal{N}(m_1, \Sigma),\ \mathcal{N}(m_2, \Sigma),$$

where m_1, m_2 are vectors of means,

$$\Sigma = \sigma \mathbf{I}$$

is a diagonal n-dimensional covariance matrix,

$$\sigma = (\sigma_1, \ldots, \sigma_n)$$

is a vector of variances. Variable X_i shall be called "noisy", if for some $i \in \{1, ..., n\}$, $\sigma_i = \sigma_{noise} >> 1$. In our experiment, the set of noisy variables

$$\{X_{i_1}, ..., X_{i_{n_{noize}}}\}$$

is chosen at random. For those variables that are not noisy, we set $\sigma_i = \sigma_0 = const$. Both classes have the same sample size.

The mixture of samples is classified by the ensemble of k-means clustering algorithms. Each algorithm performs clustering in the randomly chosen variable subspace of dimensionality n_{ens}. The ensemble decision for each pair of objects (i.e., either unite them or assign to different classes) is made by the majority voting procedure. The true overall performance P_{cor} of the ensemble is determined as the proportion of correctly classified pairs.

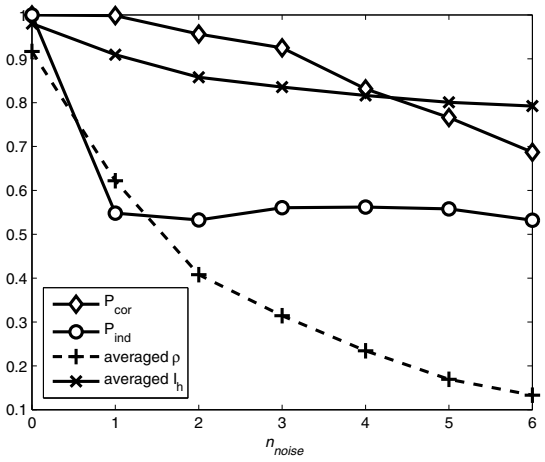

Fig. 1. Example of experiment results (averaged over 100 trials). Experiment settings: $N = 60$, $n = 10$, $m_1 = \mathbf{0}$, $m_2 = \mathbf{1}$, $\sigma_{noise} = 10$, $\sigma_0 = 0.2$, $n_{ens} = 2$, ensemble size $L = 15$.

An example of experiment results is shown in Figure 1. The graphs display the dependence of averaged values of P_{cor}, $\bar{\rho}$ and $\bar{I}_h$ from the number of noisy variables n_{noise}. To demonstrate the effectiveness of the ensemble solution in comparison with individual clustering, the averaged performance rate P_{ind} of a single k-means clustering algorithm is given (this algorithm performs clustering in the whole feature space of dimensionality n).

From this example, one can conclude that

a) in average, the ensemble algorithm has better performance than a single clustering algorithm (when a noise presents), and
b) the dynamics of estimated ensemble characteristics (averaged homogeneity and correlation) reproduces well the behavior of correct classification rate (note that this rate is directly unobserved in real clustering problems). When averaged correlation and homogeneity index are sufficiently large, one can expect good classification quality.

Conclusion

A latent variable pairwise classification model is proposed for studying non-asymptotic properties of clustering ensembles. In this model, the notions of stability, homogeneity and correlation between ensemble elements are utilized. An upper bound for probability of error is obtained. Theoretical analysis of the suggested model allows to make a conclusion that the probability of correct decision increases with an increase in number of ensemble elements. It is also found that a large degree of agreement between partial clustering solutions (expressed in our model in terms of homogeneity and correlation between ensemble elements), under condition of independence of base clustering algorithms, indicates good classification performance. Numerical experiment also confirms this conclusion.

The following possible future directions can be indicated. It is interesting to study intensional connections between the notions used in the suggested model (conditional independence, stability, homogeneity and correlation) and other known concepts such as mutual information [2] and diversity in clustering ensembles (e.g., [8,9]). Another direction could aim to improve the tightness of the obtained error bound.

Acknowledgements

This work was partially supported by the Russian Foundation for Basic Research, projects 11-07-00346a, 10-01-00113a.

References

1. Jain, A.K.: Data Clustering: 50 Years Beyond K-Means. Pattern Recognition Letters 31(8), 651–666 (2010)
2. Strehl, A., Ghosh, J.: Clustering ensembles - a knowledge reuse framework for combining multiple partitions. The Journal of Machine Learning Research 3, 583–617 (2002)

3. Kuncheva, L.I., Rodriguez, J.J., Plumpton, C.O., Linden, D.E.J., Johnston, S.J.: Random Subspace Ensembles for fMRI Classification. IEEE Transactions on Medical Imaging 29(2), 531–542 (2010)
4. Pestunov, I.A., Berikov, V.B., Kulikova, E.A.: Grid-based ensemble clustering algorithm using sequence of fixed grids. In: Proc. of the 3rd IASTED Intern. Conf. on Automation, Control, and Information Technology, pp. 103–110. ACTA Press, Calgary (2010)
5. Iam-on, N., Boongoen, T., Garrett, S.: LCE: a link-based cluster ensemble method for improved gene expression data analysis. Bioinformatics 26(12), 1513–1519 (2010)
6. Hong, Y., Kwong, S.: To combine steady-state genetic algorithm and ensemble learning for data clustering. Pattern Recognition Letters 29(9), 1416–1423 (2008)
7. Topchy, A., Law, M., Jain, A., Fred, A.: Analysis of Consensus Partition in Cluster Ensemble. In: Fourth IEEE International Conference on Data Mining, pp. 225–232. IEEE Press, New York (2004)
8. Hadjitodorov, S.T., Kuncheva, L.I., Todorova, L.P.: Moderate diversity for better cluster ensembles. Information Fusion 7(3), 264–275 (2006)
9. Azimi, J., Fern, X.: Adaptive Cluster Ensemble Selection. In: Proceedings of International Joint Conference on Artificial Intelligence, pp. 992–997 (2009)
10. Kuncheva, L.: Combining Pattern Classifiers. Methods and Algorithms. John Wiley & Sons, Hoboken (2004)
11. Breiman, L.: Random Forests. Machine Learning 45(1), 5–32 (2001)