Representative training sets for classification and the variability of empirical distributions

Saaid Baraty, Dan Simovici

University of Massachusetts Boston
sbaraty@cs.umb.edu, dsim@cs.umb.edu

Abstract. We propose a novel approach for the estimation of the size of training sets that are needed for constructing valid models in machine learning and data mining. We aim to provide a good representation of the underlying population without making any distributional assumptions.

Our technique is based on the computation of the standard deviation of the χ^2-statistics of a series of samples. When successive statistics are relatively close, we assume that the samples produced represent adequately the true underlying distribution of the population, and the models learned from these samples will behave almost as well as models learned on the entire population.

We validate our results by experiments involving classifiers of various levels of complexity and learning capabilities.

1 Introduction

Estimating a sample size that allows the inference of a good model is an important part of the learning process. We seek to determine the minimum size of a sample which is very likely to be a "fair" representative of the underlying population. Models learned from these samples will behave almost as well as models learned on the entire population and any increase in the size of the sample would result in insignificant increases in the quality of the models.

Our goal is to determine sample sizes that are sufficient to ensure that these samples adequately represent the underlying population. These samples are used as training sets for constructing models comparable in performance with those inferred from the entire population, but are cheaper to build.

Sections 2 and 3 describe in detail our approach for finding the size of a sample from a data set and a population respectively. The experimental work is presented in Section 4.

2 Estimating the size of a sample from data

Let $U = \{u_1, u_2, \ldots, u_n\}$ be a set of attributes. The set of possible states for attribute u_i, $\mathrm{Dom}(u_i)$, is assumed to be finite and is commonly referred to as *domain* of u_i. The notion of domain of attributes is extended to sets of attributes by defining the set $\mathrm{Dom}(V)$ for the set of attributes $V \subseteq U$ as $\mathrm{Dom}(V) = \prod_{v \in V} \mathrm{Dom}(v)$.

A *dataset* of U is a multi-set $\mathcal{D}$ of tuples $t \in \mathrm{Dom}(U)$. The *multiplicity* of a member t of $\mathcal{D}$ is the number $\mathcal{M}_{\mathcal{D}}(t)$ which equals the number of occurrences of tuple t in $\mathcal{D}$. The *size* of $\mathcal{D}$ is $|\mathcal{D}| = \sum_{t \in dom(U)} \mathcal{M}_{\mathcal{D}}(t)$.

Let $P_U(t)$ denote the unknown joint probability distribution of U where $t \in \mathrm{Dom}(U)$. Define the λ-*active domain* of U to be the set $\mathrm{Adom}_U(\lambda) = \{t \in \mathrm{Dom}(U) \mid P_U(t) \geq \lambda\}$, where $0 \leq \lambda < 1$ is a user specified parameter which we refer to as the *outlier threshold*.

Definition 2.1. $\mathsf{N}_{\mathcal{S}}^{\lambda} = (\mathcal{M}_{\mathcal{S}}(t_1), \ldots, \mathcal{M}_{\mathcal{S}}(t_k))$ is the *extracted frequency vector* of data sample $\mathcal{S}$ for λ. □

If $\mathsf{ELEM}(\mathcal{D}) - \overline{\mathrm{Adom}}_U^{\mathcal{D}}(\lambda) \neq \emptyset$, we add an extra tuple to account for those tuples considered as outliers, that is, we set $k = m + 1$ and $\mathcal{M}_{\mathcal{S}}(t_{m+1}) = |\mathcal{S}| - \sum_{i=1}^{m} \mathcal{M}_{\mathcal{S}}(t_i)$; otherwise we set $k = m$.

Since the tuples of sample $\mathcal{S}$ are i.i.d., we can regard the frequency vector $\mathsf{N}_{\mathcal{S}}^{\lambda}$ for an arbitrary sample $\mathcal{S}$ of fixed size q from $\mathcal{D}$ as a random vector with distribution

$$\mathsf{N}_{\mathcal{S}}^{\lambda} \sim \mathit{Multinomial}\left(q, \frac{\mathcal{M}_{\mathcal{D}}(t_1)}{|\mathcal{D}|}, \ldots, \frac{\mathcal{M}_{\mathcal{D}}(t_k)}{|\mathcal{D}|}\right), \tag{1}$$

where $q = \sum_{i=1}^{k} \mathcal{M}_{\mathcal{S}}(t_i)$ and $\mathcal{M}_{\mathcal{D}}(t_k) = |\mathcal{D}| - \sum_{i=1}^{k-1} \mathcal{M}_{\mathcal{D}}(t_i)$.

Define the χ^2-*statistics of sample* $\mathcal{S}$ *for outlier threshold* λ *with respect to target probability distribution* $\mathbf{p} = (p_1, \ldots, p_k)$ as $\mathcal{X}^2_{\mathcal{S}}(\lambda, \mathbf{p}) = \sum_{i=1}^{k} \frac{(\mathcal{M}_{\mathcal{S}}(t_i) - q p_i)^2}{q p_i}$. We refer to $\mathcal{X}^2_{\mathcal{S}}(\lambda, \mathbf{p})$ as χ^2-statistics, because if $\mathsf{N}_{\mathcal{S}}^{\lambda} \sim \mathit{Multinomial}\,(q, p_1, \ldots, p_k)$ then, as $q \to \infty$, the distribution of the random variable $\mathcal{X}^2_{\mathcal{S}}(\lambda, \mathbf{p})$ converges in distribution to χ^2-distribution with $k - 1$ degree of freedom (Pearson, 1900). We use $\mathcal{X}^2_{\mathcal{S}}(\lambda, \mathbf{p})$ as a measure of how close $\mathsf{N}_{\mathcal{S}}^{\lambda}$ is in representing the target distribution $\mathbf{p}$. As we increase the sample size q, by the strong law of large numbers, $\mathcal{X}^2_{\mathcal{S}}(\lambda, \mathbf{p})$ becomes smaller.

Our aim is to estimate q, the size of a sample from data, such that the extracted frequency vectors of the samples of size q are likely to closely represent the empirical distribution of $\mathcal{D}$ for those tuples that are not λ-outliers. Therefore, we specify the target distribution to be the empirical distribution of the data and define χ^2-*statistics of data sample* $\mathcal{S}$ *for outlier threshold* λ *with respect to empirical distribution of data set* $\mathcal{D}$ to be

$$\mathcal{X}^2_{\mathcal{S}}(\lambda, \mathcal{D}) = \sum_{i=1}^{k} \frac{\left(\mathcal{M}_{\mathcal{S}}(t_i) - q\frac{\mathcal{M}_{\mathcal{D}}(t_i)}{|\mathcal{D}|}\right)^2}{q\frac{\mathcal{M}_{\mathcal{D}}(t_i)}{|\mathcal{D}|}} = \frac{|\mathcal{D}|}{q}\sum_{i=1}^{k} \frac{\mathcal{M}^2_{\mathcal{S}}(t_i)}{\mathcal{M}_{\mathcal{D}}(t_i)} - q.$$

Let $\mathcal{S}_1, \ldots, \mathcal{S}_z$ be repeatedly drawn z samples of a fixed size q from $\mathcal{D}$. Given a threshold λ we compute $\mathcal{X}^2_{\mathcal{S}_i}(\lambda, \mathcal{D})$ for each $\mathcal{S}_i$ followed by This process is summarized in Algorithm 1, where $\hat{\sigma}_q$ is the standard deviation among values $\mathcal{X}^2_{\mathcal{S}_i}(\lambda, \mathcal{D})$ for different i.

3 An iterative estimation of the size of a sample from a population

We apply our approach to estimate the size of a fair sample from a population without having a data set at hand. Since $P_U(t)$ is unknown, we assume that $\mathrm{Adom}_U(\lambda) = \{t_1, \ldots, t_m\}$ for some outlier threshold λ.

Algorithm 1: The pseudocode for finding the size of a sufficient training set from data set $\mathcal{D}$.

foreach *sample size q from smallest to largest* **do**
  draw data samples of size q: $\mathcal{S}_1, \ldots, \mathcal{S}_z$ with replacement from data set $\mathcal{D}$;
  compute the standard deviation $\hat{\sigma}_q$ of sequence $\mathcal{X}^2_{\mathcal{S}_1}(\lambda, \mathcal{D}), \ldots, \mathcal{X}^2_{\mathcal{S}_z}(\lambda, \mathcal{D})$;
output: sample size q such that for any sample size $v \geq q$ we have $\hat{\sigma}_q \approx \hat{\sigma}_v$

Definition 3.1. The *extracted frequency vector of a population sample* $\mathcal{S}$ for outlier threshold λ is $\mathsf{M}^\lambda_\mathcal{S} = (\mathcal{M}_\mathcal{S}(t_1), \ldots, \mathcal{M}_\mathcal{S}(t_k))$ where, as in Definition 2.1, we have two cases: (1) if $\mathrm{Dom}(U) - \mathrm{Adom}_U(\lambda) \neq \emptyset$ we add an extra tuple to account for those tuples considered as outliers, that is, we set $k = m + 1$ and $\mathcal{M}_\mathcal{S}(t_{m+1}) = |\mathcal{S}| - \sum_{i=1}^{m} \mathcal{M}_\mathcal{S}(t_i)$, and (2) otherwise, that is, if $\mathrm{Dom}(U) = \mathrm{Adom}_U(\lambda)$ we set $k = m$. □

Informally, we consider a sample of size q as a λ-fair representative of the population if $\mathsf{M}^\lambda_\mathcal{S}/q$ closely approximates the population's true distribution vector of the tuples in $\mathrm{Adom}_U(\lambda)$. Similar to previous section, we treat $\mathsf{M}^\lambda_\mathcal{S}$ for arbitrary population sample $\mathcal{S}$ of size q as a random vector $\mathsf{M}^\lambda_\mathcal{S} \sim \mathit{Multinomial}\,(q, P_U(t_1), \ldots, P_U(t_k))$, where $P_U(t_k) = 1 - \sum_{i=1}^{k-1} P_U(t_i)$. However, the probabilities $P_U(t_i)$ for $1 \leq i \leq k$ are unknown. Hence, we define the random probability vector $\mathbf{p} = (p_1, \ldots, p_k)$ to represent the occurrence of a k-dimensional probability distribution as the true underlying distribution of the population. Then, $\mathbf{p}$ is represented by the probability space $(\Omega, \mathcal{P}(\Omega), f)$ of k-dimensional probability distribution vectors where the sample space Ω is a standard $(k-1)$-simplex.

Note that $\mathcal{X}^2_\mathcal{S}(\lambda, \mathbf{p})$ is a random variable itself with values in $\mathbb{R}_{\geq 0}$. We approximate the χ^2-statistics of a population sample $\mathcal{S}$ with respect to the true underlying distribution by the conditional expected value of $\mathcal{X}^2_\mathcal{S}(\lambda, \mathbf{p})$ given that we have another sample of the same size from the same population at hand. This conditioned sample approximates the shape of the probability distribution of $\mathbf{p}$ (the second order distribution) if it is large enough to unbiasedly represent the underlying distribution of the population. Let $\mathcal{S}_1, \ldots, \mathcal{S}_{2z}$ be a sequence of independent samples of size q drawn uniformly at random with replacement from the underlying population.

We compute the conditional expected value of χ^2-statistics (CECS statistics) $E[\mathcal{X}^2_{\mathcal{S}_i}(\lambda, \mathbf{p})|\mathcal{S}_{z+i}]$ for $1 \leq i \leq z$. This statistics is used as a substitute for the actual χ^2-statistics of $\mathcal{S}_i$ with respect to target distribution P_U. Next, we we compute the standard deviation among the CECS statistics of $\mathcal{S}_i$ given $\mathcal{S}_{z+i}$ for $1 \leq i \leq z$. If q is large enough, then the probability distributions captured by the frequencies extracted from $\mathcal{S}_i$ and $\mathcal{S}_{z+i}$ would be similar to P_U and thus, similar to each other. Therefore, the variation in CECS statistics is expected to be small. Next, observe that $P(\mathcal{S}_\ell|\mathbf{p}) = \prod_{j=1}^{k} p_j^{\mathcal{M}_{\mathcal{S}_\ell}(t_j)}$. If we further assume the prior, $\mathbf{p} \sim \mathit{Dirichlet}(\mu_1, \ldots, \mu_k)$ for $\mu_1, \ldots, \mu_k > 0$, then, it can be shown that $f(\mathbf{p}|\mathcal{S}_\ell)$ follows the Dirichlet distribution, $\mathit{Dirichlet}(\alpha_1, \ldots, \alpha_k)$ of order $k \geq 2$.

We draw (with replacement) simple random samples $\mathcal{S}_1, \ldots, \mathcal{S}_{2z}$ of size q from $\mathcal{OP}$, where $|\mathcal{OP}|$ is a domain-dependent multiple of q and the larger the size of the observation pool is relative to q, the more reliable is the conclusion of the process. $E[\mathcal{X}^2_{\mathcal{S}_i}(\lambda, \mathbf{p})|\mathcal{S}_{z+i}]$ is evaluated for each i. If the standard deviation among conditional expectations is sufficiently small and stabilizes at a certain value of q, then we choose this value of q as the threshold of the size

of fair samples or adequate training/evaluation sets. Otherwise, we increase q and repeat the process.

In this iterative process we may need to expand the observation pool to make sure it is a substantial multiple of q. As we add new observations to our pool, we need to update $\overline{\mathrm{Adom}}_U^{\mathcal{OP}}(\lambda)$, the set of tuples to be considered according to outlier threshold λ, and subsequently k, the number of dimensions of the probability space. Observe that as we expand the observation pool $\mathcal{OP}$, $\overline{\mathrm{Adom}}_U^{\mathcal{OP}}(\lambda)$ becomes a closer approximation of the set $\mathrm{Adom}_U(\lambda)$. The following pseudocode explains the process of finding the size of a fair sample from a population as explained in this section.

Algorithm 2: The pseudocode for finding size of a sufficient training set from a population.

foreach *sample size q from smallest to largest* **do**
  if $\neg(|\mathcal{OP}| >> q)$ **then**
    expand the $\mathcal{OP}$ such that $|\mathcal{OP}| \gg q$;
    evaluate $\overline{\mathrm{Adom}}_U^{\mathcal{OP}}(\lambda)$ and find k based on this set;
  draw independent samples of size q: $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_{2z}$ (with replacement) from $\mathcal{OP}$;
  compute the standard deviation $\hat{\sigma}_q$ of sequence
  $E[\mathcal{X}^2_{\mathcal{S}_1}(\lambda, \mathbf{p})|\mathcal{T}_1], \ldots, E[\mathcal{X}^2_{\mathcal{S}_z}(\lambda, \mathbf{p})|\mathcal{T}_z]$;

output: sample size q such that for any size $v \geq q$ we have $\hat{\sigma}_q \approx \hat{\sigma}_v$

4 Experimental results

In the first experiment we employed the Algorithm 1 to estimate the size of a data sample from the Bank Marketing Data Set (Moro et al., 2011) which contains $45,211$ records (see Figure 1). For $\lambda = 0$, the standard deviation drops to its minimal level when q is around $5,000$ so a sample of size $5,000$ is very likely to fairly represent the entire data which is of size $45,211$. For $\lambda = 0.00039$ a training sample of size $2,000$ is suitable.

In the next experiment we evaluated our approach for determining the size of a representative sample from a population as summarized in Algorithm 2 for $\epsilon = 0.005$. We simulated the process of gathering observations from a population in order to expand the observation pool by synthetically generating tuples of four attributes using a multinomial distribution with randomly selected parameters, $|\mathrm{Dom}(U)| = 24$ and $\lambda = 0$ and we executed the Algorithm 2 with $z = 1000$.

For each q we generated one hundred samples of size q from a synthetic data set and used WEKA to learn a k-nearest neighbor (k-NN) classifier from each sample. Then, we evaluated the prediction performance of the classifiers using a fixed test set of size $10,000$ which is large enough to represent unbiasedly the underlying distribution of the domain. The average and standard deviation of the percentage of correctly classified instances (CCI) are shown in Figure 2. Similar results were obtained for Bayesian Networks.

On the other hand, experiments with naive Bayes classifiers yield quite different results shown in Figure 3. The improvement in average percentage of CCI as a result of increasing

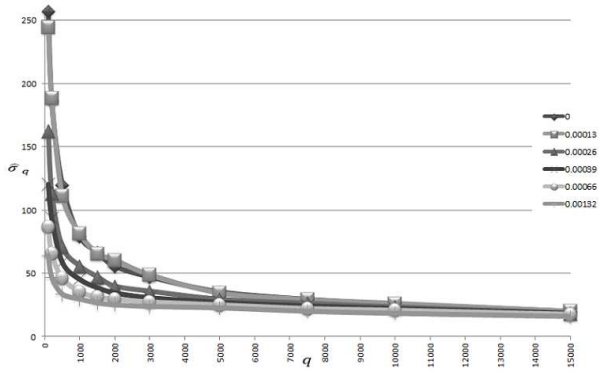

FIG. 1 – *Standard deviation of χ^2-statistics of data samples with respect to changes in sample size q for Bank Marketing data. Each curve corresponds to a particular value of outlier threshold λ listed in the right hand side.*

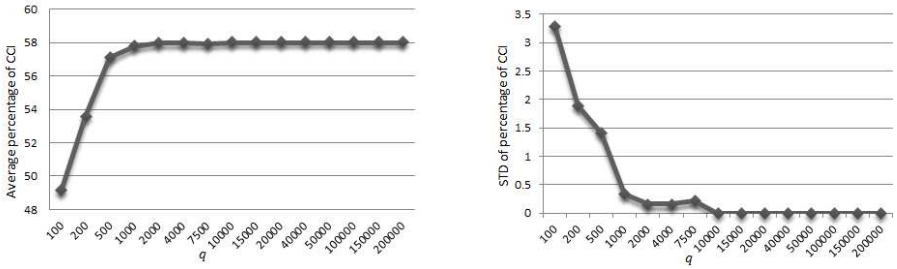

FIG. 2 – *Average and STD for* 20*-NN*

the sample size q is much smaller than in the previous cases and the average percentage of CCI reaches its peak at sample size $q = 2,000$ and then slightly decreases to a constant level afterwards. Finally, the standard deviation of the percentage of CCI converges to zero slower than previous two cases. These differences are due to the fact that naive Bayes classifiers are less dependent on the global joint probability distribution than k-NN classifiers and Bayesian networks because of the naive independence assumption.

The experimental results show that it does not make sense to go beyond the size that we determine here, because the improvement we gain in the performance is insignificant or inexistent. If the evaluated size of the training set is prohibitively large, then, we may be able to reduce the sample size approximation by analyzing it in the context of a specific classifier.

References

Pearson, K. (1900), On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to

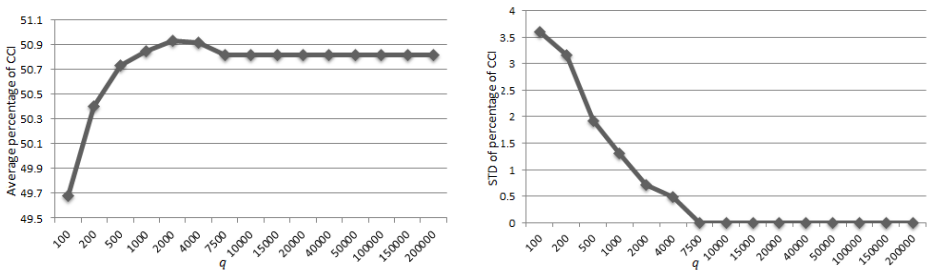

FIG. 3 – *The average and standard deviation of the percentage of correctly classified instances for naive Bayes classifiers*

have arisen from random sampling, Philosophical Magazine, Vol. 50, no. 302, ser. 5, pp. 157–175.

Moro, S., Laureano, R. and Cortez, P. (2005), Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology, Proceedings of the European Simulation and Modelling Conference, pp. 117-121. Portugal.

Schmidtmann, I., Hammer, G., Sariyar, M. and Gerhold-Ay, A. (2009), Evaluation des Krebsregisters NRW Schwerpunkt Record Linkage, Technical Report, IMBEI

Résumé

Nous proposons une nouvelle approche pour l'estimation de la taille des ensembles d'apprentissage qui sont nécessaires pour construire des modèles valides dans l'extraction de connaissaices. Nous visons à fournir une bonne représentation de l'ensemble de données sans faire des hypothèses de répartition.

Notre technique est basée sur le calcul de l'écart-type des χ^2-statistiques d'une série d'échantillons. Lorsque les statistiques successives sont relativement proches, nous supposons que les échantillons produits représentent adéquatement la vraie distribution sous-jacente de la population, et les modèles tirés de ces échantillons se comportent presque aussi bien que les modèles appris sur l'ensemble de la population.

Nous validons nos résultats par des travaux expérimentaux impliquant une variété des classificateurs.