Bounding the Difference Between RankRC and RankSVM and Application to Multi-Level Rare Class Kernel Ranking

Aditya Tayal[a,1,*], Thomas F. Coleman[b,1,2], Yuying Li[a,1]

[a] *Cheriton School of Computer Science, University of Waterloo, Waterloo, ON, Canada N2L 3G1*
[b] *Combinatorics and Optimization, University of Waterloo, Waterloo, ON, Canada N2L 3G1*

Abstract

Rapid explosion in data accumulation has yielded larger and larger data mining problems. Many practical problems have intrinsically unbalanced or rare class distributions. Standard classification algorithms, which focus on overall classification accuracy, often perform poorly in these cases. Recently, Tayal et al. (2013) proposed a kernel method called RankRC for large-scale unbalanced learning. RankRC uses a ranking loss to overcome biases inherent in standard classification based loss functions, while achieving computational efficiency by enforcing a rare class hypothesis representation. In this paper we establish a theoretical bound for RankRC by establishing equivalence between instantiating a hypothesis using a subset of training points and instantiating a hypothesis using the full training set but with the feature mapping equal to the orthogonal projection of the original mapping. This bound suggests that it is optimal to select points from the rare class first when choosing the subset of data points for a hypothesis representation. In addition, we show that for an arbitrary loss function, the Nyström kernel matrix approximation is equivalent to instantiating a hypothesis using a subset of data points. Consequently, a theoretical bound for the Nyström kernel SVM can be established based on the perturbation analysis of the orthogonal projection in the feature mapping. This generally leads to a tighter bound in comparison to perturbation analysis based on kernel matrix approximation. To further illustrate computational effectiveness of RankRC, we apply a multi-level rare class kernel ranking method to the Heritage Health Provider Network's health prize competition problem and compare the performance of RankRC to other existing methods.

Keywords: rare class, receiver operator characteristic, AUC, ranking loss, scalable computing, nonlinear kernel, ranking svm

[*] Corresponding author

Email addresses: amtayal@uwaterloo.ca (Aditya Tayal), tfcoleman@uwaterloo.ca (Thomas F. Coleman), yuying@uwaterloo.ca (Yuying Li)

[1] All three authors acknowledge funding from the National Sciences and Engineering Research Council of Canada

[2] This author acknowledges funding from the Ophelia Lazaridis University Research Chair. The views expressed herein are solely from the authors.

1. Introduction

Rapid data accumulation has yielded larger and larger data mining problems. Many practical problems naturally arise as rare class problems. Applications of the rare class prediction include fraud detection, customer churn, intrusion detection, fault detection, credit default, insurance risk, and health management. The rare class prediction problem is also referred to as an unbalanced or skewed class distribution problem (He and Garcia, 2009). In these problems samples from one class are extremely rare (the minority class), while samples belonging to the other class(es) are plenty (the majority class). Standard classification methods, which include support vector machines (SVM) (Japkowicz and Stephen, 2002; Raskutti and Kowalczyk, 2004; Wu and Chang, 2003), decision trees (Batista et al., 2004; Chawla et al., 2004; Japkowicz and Stephen, 2002; Weiss, 2004), neural networks (Japkowicz and Stephen, 2002), Bayesian networks (Ezawa et al., 1996), and nearest neighbor methods (Batista et al., 2004; Zhang and Mani, 2003), perform poorly when dealing with unbalanced data. This is because they attempt to minimize total classification error. However, in rare class problems, minority examples constitute a small proportion of the data and have little impact on the total error. Thus majority examples overshadow the minority class, resulting in models that are heavily biased in recognizing the majority class. Also, errors from different classes are assumed to have the same costs, which is usually not true in practice. In most problems, correct classification of the rare class is more important.

Solutions to the class imbalance problem have been proposed at both the data and algorithm levels. At the data level, various resampling techniques are used to balance class distribution, including random under-sampling of majority class instances (Kubat and Matwin, 1997), over-sampling minority class instances with new synthetic data generation (Chawla et al., 2002), and focused resampling, in which samples are chosen based on additional criteria (Zhang and Mani, 2003). Although sampling approaches have achieved success in some applications, they are known to have drawbacks. For instance under-sampling can eliminate useful information, while over-sampling can result in overfitting. At the algorithm level, solutions are proposed by adjusting the algorithm itself. This usually involves adjusting the costs of the classes to counter the class imbalance (Turney, 2000; Lin et al., 2000; Chang and Lin, 2011) or adjusting the decision threshold (Karakoulas and Shawe-Taylor, 1999). However, true error costs are often unknown and using an inaccurate cost model can lead to additional bias.

A data mining method, either explicitly or implicitly, has three key components. Firstly, an

empirical loss function is minimized. Secondly, model complexity is minimized, for example via regularization. Thirdly, a mechanism balances the tradeoff between these two objectives. Choosing an appropriate loss function is important and should take into account potential class imbalances. In addition, scalability and computational efficiency become critical factors as data mining problems continually grow in size.

In this paper, we focus on kernel based methods. We use an appropriate loss function for rare class problems and exploit class imbalance to achieve an efficient space and time algorithm, making it feasible for large-scale problems. Rare class problems can be viewed as consisting of dual, conflicting objectives: (1) accuracy of the minority class or true positive rate and (2) inaccuracy of the minority class or false positive rate. The Receiver Operator Characteristic (ROC) curve graphically depicts these criteria in a two-dimensional domain, where each axis represents one objective. The area under the ROC curve (AUC) summarizes the curve in a single numerical measure and is often used as an evaluation metric for unbalanced problems (He and Garcia, 2009; Bradley, 1997). We use the AUC to define an appropriate loss function for rare class problems. Maximizing AUC is equivalent to minimizing a biclass ranking loss. A convex approximation of the biclass ranking loss leads to a RankSVM (Herbrich et al., 2000) problem with two ordinal levels. However, solving the dual optimization problem to obtain a kernel RankSVM solution, requires $O(m^6)$ time and $O(m^4)$ space, where m is the number of data samples. Chapelle and Keerthi (2010) propose a primal approach to solve RankSVM, which results in $O(m^3)$ time and $O(m^2)$ space for nonlinear kernels. However, this is still computationally prohibitive for large data mining problems.

Recently, Tayal et al. (2013) propose a rare class based kernel method for unbalanced problems called RankRC. Like RankSVM, RankRC uses a ranking loss that maximizes AUC for two ordinal levels and is suitable for unbalanced problems. In addition, RankRC uses a rare class hypothesis representation to achieve significant computational advantage. RankRC can be solved in $O(mm_+)$ time and $O(mm_+)$ space, where m_+ is the number of rare class examples. Computational results in Tayal et al. (2013) demonstrate that RankRC performs similar to kernel RankSVM for rare class problems, while able to scale to much larger datasets.

The main objective of this paper is to theoretically establish the difference between RankRC, in which the hypothesis is instantiated by rare class points, and RankSVM, in which a hypothesis is instantiated by the full data set. We also extend RankRC to multi-level rare class ranking problems.

Specifically, the contributions of this paper are as follows:

- We mathematically establish an upper bound on the difference between the optimal hypotheses of RankSVM and RankRC. This bound is established by observing that a regularized loss minimization problem with a hypothesis instantiated with points in a subset is equivalent to a regularized loss minimization problem with a hypothesis instantiated by the full data set but using the orthogonal projection of the original feature mapping.

- We show that the upper bound suggests that, under the assumption that a hypothesis is instantiated by a subset of data points of a fixed cardinality, it is optimal to choose data points from the rare class first.

- We further demonstrate that the Nyström kernel approximation method is equivalent to solving a kernel regularized loss problem instantiated by a subset of data points corresponding to the selected columns. Consequently, a theoretical bound for Nyström kernel approximation methods can be established based on the perturbation analysis of the orthogonal projection in the feature mapping. We demonstrate that this can provide a tighter bound in comparison to perturbation analysis based on kernel matrix approximation, which can be arbitrarily large depending on the condition number of the approximation matrix.

- We extend the biclass RankRC formulation to multi-level ranking and apply it to a recent competition problem sponsored by the Heritage Health Provider Competition. The problem illustrates how RankRC can be used for ordinal regression where one ordinal level contains the vast majority of examples. We compare performance of RankRC with other methods and demonstrate computational and predictive advantages.

The rest of the paper is organized as follows. Section 2 reviews the AUC measure, RankSVM, RankRC, and its extension to multiple ordinal levels. Section 3 presents theoretical justification for RankRC by comparing to RankSVM. Section 4 discusses connection with the Nyström approximation method. Section 5 describes the computational results from applying RankRC to the HPN hospital admission prediction problem. Finally, Section 6 concludes with summary remarks and potential extensions.

2. AUC, Ranking Loss, and Rare Class Ranking

Evaluation metrics play an important role in learning algorithms. They provide ways to assess performance as well as guide modeling. For classification problems, error rate is the most commonly used assessment metric. Consider the two-class case first. Assume that $\mathcal{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), ..., (\mathbf{x}_m, y_m)\}$ is a set of m training examples, where $\mathbf{x}_i \in \mathcal{X} \subseteq \mathbb{R}^d$, $y_i \in \{+1, -1\}$. Then the empirical error rate for an inductive hypothesis, $f(\mathbf{x})$, typically obtained by training on example set $\mathcal{D}$, is defined as,

$$\text{Error Rate} = \frac{1}{m} \sum_{i=1}^{m} \mathbb{I}[f(\mathbf{x}_i) \neq y_i], \tag{1}$$

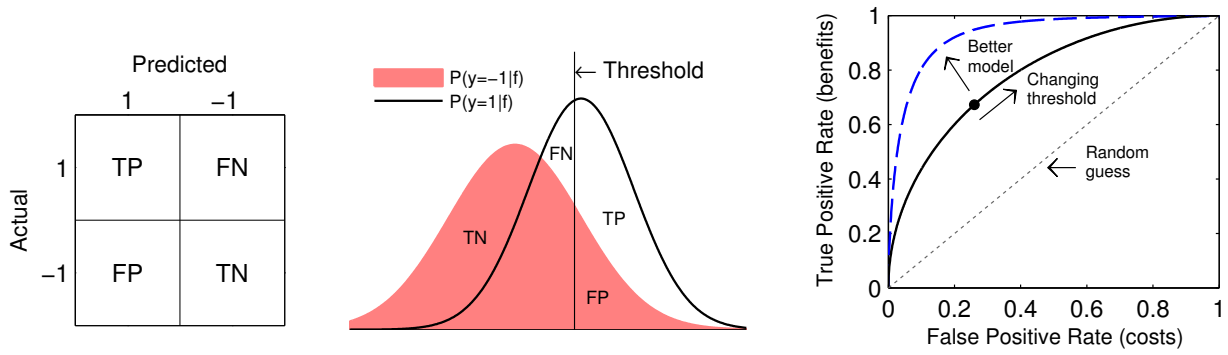
where $\mathbb{I}[p]$ denotes the indicator function and is equal to 1 if p is true, 0 if p is false. However, for highly unbalanced datasets, this error rate is not appropriate since it can be biased toward the majority class (Provost et al., 1997; Maloof, 2003; Sun et al., 2007; He and Garcia, 2009). For example, if a given dataset includes 1 percent of positive class examples and 99 percent of negative examples, a naive solution which assigns every example to be positive will obtain only 1 percent error rate. Indeed, classifiers that always predict the majority class can obtain lower error rates than those that predict both classes equally well. But clearly these are not useful hypotheses.

2.1. Maximizing AUC and Minimizing Ranking Loss

When the class distribution is unbalanced, classification performance is more accurately represented by a confusion matrix. For binary classification problems, a 2-by-2 confusion matrix with rows corresponding to actual targets and columns corresponding to the predicted values can be used (see Figure 1a). The off-diagonal entries, denoting false negatives and false positives, together represent the total number of errors. A single performance measure that uses values from both rows, e.g., the error rate in (1), will be sensitive to the class skew.

In this paper, we follow convention and set the minority class as positive and the majority class as negative, with m_+ denoting the number of minority examples and m_- the number of majority ones.

The Receiver Operating Characteristic (ROC) can be used to obtain a skew independent measure (Provost et al., 1997; Bradley, 1997; Metz, 1978). Most classifiers intrinsically output a numerical score and a predicted label is obtained by thresholding this score. For example, a threshold of zero leads to taking the sign of the numerical output as the label. Each threshold value generates



(a) (b) (c)

Figure 1: ROC analysis. (a) Shows 2×2 confusion matrix representing the results of a model. TP, FP, FN, and TN, denote True Positives, False Positives, False Negatives and True Negatives, respectively. (b) Different quantities of TP, FP, FN and TN are obtained as the threshold value of a model is adjusted. (c) The ROC curve plots TP rate against FP rate for different threshold values. The dashed (blue) ROC curve dominates the solid (black) ROC curve. The dotted (gray) ROC curve has an AUC of 0.5, indicating a model with no discriminative value.

a confusion matrix with different quantities of false positives and negatives (see Figure 1b). The ROC graph is obtained by plotting the true positive rate (number of true positives divided by m_+) against the false positive rate (number of false positives divided by m_-) as the threshold level is varied (see Figure 1c). It depicts the trade-off between benefits (true positives) and costs (false positives) for different choices of the threshold. Thus it does not depend on a priori knowledge or specification of the cost context or the class distribution. A ROC curve that dominates another provides a better solution at any cost point.

To facilitate comparison, it is convenient to characterize ROC curves using a single measure. The area under a ROC curve (AUC) can be used for this purpose. It is the average performance of the model across all threshold levels and corresponds to the Wilcoxon rank statistic (Hanley and Mcneil, 1982). AUC represents the probability that the score generated by a classifier places a positive class sample above a negative class sample when the positive sample is randomly drawn from the positive class and the negative sample is randomly drawn from the negative class (DeLong et al., 1988). The AUC can be computed by forming the ROC curve and using the trapezoid rule to calculate the area under the curve. Also, given the intrinsic output of a hypothesis, $f(\mathbf{x})$, we can directly compute the empirical AUC by counting pairwise correct rankings (DeLong et al., 1988):

$$\text{AUC} = \frac{1}{m_+ m_-} \sum_{\{i:y_i=+1\}} \sum_{\{j:y_j=-1\}} \mathbb{I}\left(f(\mathbf{x}_i) \geq f(\mathbf{x}_j)\right) \ . \tag{2}$$

Instead of maximizing accuracy (minimizing error rate) in the modeling problem, we maxi-

mize AUC (minimize 1-AUC) as an alternative loss function, which is more appropriate for unbalanced datasets. In practice, $\mathbb{I}[-p]$ is often replaced with a convex approximation such as the hinge, logistic, or exponential cost functions (Bartlett et al., 2006). Specifically, using the hinge function, $\ell_h(p) = \max(0, 1-p)$, and controlling model complexity with ℓ_2-regularization, leads to the following convex ranking problem,

$$\min_{\mathbf{w}\in\mathbb{R}^d} \quad \frac{1}{m_+ m_-} \sum_{\{i:y_i=+1\}} \sum_{\{j:y_j=-1\}} \ell_h\left(\mathbf{w}^T\mathbf{x}_i - \mathbf{w}^T\mathbf{x}_j\right) + \frac{\lambda}{2}\|\mathbf{w}\|_2^2 \,, \tag{3}$$

where $\lambda \in \mathbb{R}_+$ is a penalty parameter for model complexity. Here, the hypothesis, $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$, is assumed linear in the input space $\mathcal{X}$. Problem (3) is a special case of RankSVM proposed by Herbrich et al. (2000) with two ordinal levels. Like the standard SVM, RankSVM also leads to a dual problem which can be expressed in terms of dot-products between input vectors. This allows us to obtain a non-linear function through the kernel trick (Boser et al., 1992), which consists of using a kernel, $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, that corresponds to a feature map, $\phi : \mathcal{X} \to \mathcal{F} \subseteq \mathbb{R}^{d'}$, such that $\forall \mathbf{u}, \mathbf{v} \in \mathcal{X}$, $k(\mathbf{u}, \mathbf{v}) = \phi(\mathbf{u})^T\phi(\mathbf{v})$. The kernel k directly computes the inner product of two vectors in a potentially high-dimensional feature space $\mathcal{F}$, without the need to explicitly form the mapping. Consequently, we can replace all occurrences of the dot-product with k in the dual and work implicitly in space $\mathcal{F}$.

Since there is a Lagrange multiplier for each constraint associated with the hinge loss, the dual formulation leads to a problem in $m_+m_- = O(m^2)$ variables. Assuming the optimization procedure has cubic complexity in the number of variables and quadratic space requirements, the complexity of the dual method becomes $O(m^6)$ time and $O(m^4)$ space, which is unreasonably large for even medium sized datasets.

As noted by Chapelle (2007) and Chapelle and Keerthi (2010), we can also solve the primal problem in the implicit feature space due to the Representer Theorem (Kimeldorf and Wahba, 1970; Schölkopf et al., 2001). This theorem states that the solution of any regularized loss minimization problem in $\mathcal{F}$ can be expressed as a linear combination of kernel functions evaluated at the training samples, $k(\mathbf{x}_i, \cdot)$, $i = 1, \ldots, m$. Thus, the solution of (3) in $\mathcal{F}$ can be written as:

$$f(\mathbf{x}) = \sum_{i=1}^{m} \beta_i k(\mathbf{x}_i, \mathbf{x}) \,, \text{ or } \mathbf{w} = \sum_{i=1}^{m} \beta_i k(\mathbf{x}_i, \cdot) \,. \tag{4}$$

Substituting (4) in (3) we can express the primal problem in terms of $\boldsymbol{\beta}$,

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^m} \quad \frac{1}{m_+ m_-} \sum_{\{i:y_i=+1\}} \sum_{\{j:y_j=-1\}} \ell_h\left(K_{i\bullet}\boldsymbol{\beta} - K_{j\bullet}\boldsymbol{\beta}\right) + \frac{\lambda}{2}\boldsymbol{\beta}^T K \boldsymbol{\beta}\,, \tag{5}$$

where $K \in \mathbb{R}^{m\times m}$ is a positive semi-definite kernel matrix, $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$, and $K_{i\bullet}$ denotes the ith row of K. To be able to solve (5) using unconstrained optimization methods, we require the objective to be differentiable. We replace the hinge loss, ℓ_h, with an ϵ-smoothed differentiable approximation, ℓ_ϵ, defined as,

$$\ell_\epsilon(z) = \begin{cases} (1-\epsilon) - z & \text{if } z < 1-2\epsilon \\ \frac{1}{4\epsilon}(1-z)^2 & \text{if } 1-2\epsilon \le z < 1 \\ 0 & \text{if } z \ge 1\,, \end{cases}$$

which transitions from linear cost to zero cost using a quadratic segment of length 2ϵ. We note that ℓ_ϵ provides similar benefits as the hinge loss (Rosset et al., 2003; Nguyen et al., 2009). Thus we can solve (5) using standard unconstrained optimization techniques. Since there are m variables, Newton's method would, for example, take $O(m^3)$ operations per iteration.

RankSVM is popular in the information retrieval community, where linear models are the norm (e.g. see Joachims, 2002). For a linear model, with d-dimension input vectors, the complexity of RankSVM can be reduced to $O(md + m\log m)$ (Chapelle and Keerthi, 2010). However, many rare class problems require a nonlinear function to achieve optimal results. Solving a nonlinear RankSVM requires $O(m^3)$ time and $O(m^2)$ space (Chapelle and Keerthi, 2010), which is still not practical for mid- to large-sized datasets. We believe this complexity is, in part, the reason why nonlinear RankSVMs are not commonly used to solve rare class problems.

2.2. RankRC: ranking with a rare class representation

To make RankSVM computationally feasible for large scale unbalanced problems, Tayal et al. (2013) propose to use a hypothesis instantiated by rare class points only. We present motivation for RankRC by assuming specific properties of the class conditional distributions and kernel function. Zhu et al. (2006) make use of similar assumptions, however, in their method they attempt to directly estimate the likelihood ratio. In contrast, RankRC uses a regularized loss minimization approach.

The optimal ranking function for a binary classification problem is the posterior probability,

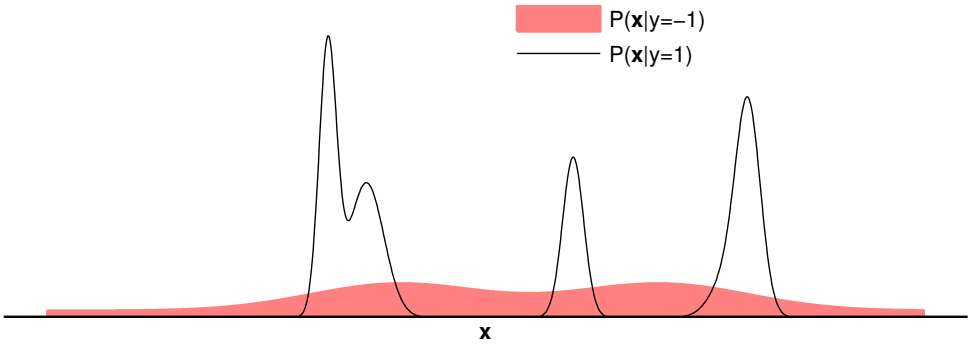


Figure 2: Example class conditional distributions for a rare class dataset showing that $P(\mathbf{x}|y=1)$ is concentrated with bounded support, while $P(\mathbf{x}|y=-1)$ is relatively constant in the local regions around the positive class.

$P(y=1|\mathbf{x})$, since it minimizes Bayes risk for arbitrary costs. From Bayes' Theorem, we have,

$$P(y=1|\mathbf{x}) = \frac{P(y=1)P(\mathbf{x}|y=1)}{P(y=1)P(\mathbf{x}|y=1)+P(y=-1)P(\mathbf{x}|y=-1)} \ . \tag{6}$$

Also, any monotonic transformation of (6) yields equivalent ranking capability. Dividing both the numerator and denominator of (6) by $P(y=-1)P(\mathbf{x}|y=-1)$, it can be observed that $P(y=1|\mathbf{x})$ is a monotonic transformation of the likelihood ratio

$$f(\mathbf{x}) = \frac{P(\mathbf{x}|y=1)}{P(\mathbf{x}|y=-1)} \ . \tag{7}$$

Using kernel density estimation, the conditional density $P(\mathbf{x}|y=1)$ can be approximated using

$$P(\mathbf{x}|y=1) = \frac{1}{m_+} \sum_{\{i:y_i=+1\}} k(\mathbf{x},\mathbf{x}_i;\sigma) \ , \tag{8}$$

where $k(\mathbf{x},\mathbf{x}_i;\sigma)$ is a kernel density function, typically a smooth unimodal function of $\mathbf{x}$ with a peak at $\mathbf{x}_i$ and a width localization parameter $\sigma > 0$. This kernel density estimation encompasses a large range of possible distributions from the m_+ rare examples provided.

In rare class problems, most examples are from the majority class ($y=-1$) and only a small number of samples are from the rare class ($y=1$). It is reasonable to assume that the minority class examples are concentrated in local regions with bounded support, while the majority class acts as background noise. Therefore, in a neighborhood around the minority class examples, the conditional density function $P(\mathbf{x}|y=-1)$ can be assumed to be relatively flat in comparison to $P(\mathbf{x}|y=1)$, see Figure 2 for instance. Assume $P(\mathbf{x}|y=-1) \approx c_i$ for each minority example i in the neighborhood of $\mathbf{x}_i$. Together with (8), the likelihood ratio (7) can be written in the form below,

$$f(\mathbf{x}) = \sum_{\{i:y_i=+1\}} \beta_i k(\mathbf{x}_i, \mathbf{x}; \sigma) \,, \tag{9}$$

which uses only kernel function evaluations of the minority class. In contrast to (4), this formulation takes specific advantage of the conditional density structure of rare class problems.

Replacing the kernel density function with a general kernel function and substituting (9) in (3) results in the following RankRC problem,

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^{m_+}} \quad \frac{1}{m_+ m_-} \sum_{\{i:y_i=+1\}} \sum_{\{j:y_j=-1\}} \ell_h \left(K_{i+}\boldsymbol{\beta} - K_{j+}\boldsymbol{\beta}\right) + \frac{\lambda}{2}\boldsymbol{\beta}^T K_{++}\boldsymbol{\beta} \,, \tag{10}$$

where K_{i+} denotes ith row of K with column entries corresponding to the positive class, and $K_{++} \in \mathbb{R}^{m_+ \times m_+}$ is the square submatrix of K corresponding to positive class entries. By replacing ℓ_h with a smooth differentiable loss, ℓ_ϵ, problem (10) can be solved in $O(mm_+)$ time and $O(mm_+)$ space (see Tayal et al., 2013, for detailed discussion). Based on several synthetic and real rare class problems, it is shown in Tayal et al. (2013) that RankRC is computationally more efficient and can scale to large datasets, while not sacrificing test performance compared to RankSVM.

3. Theoretical Comparison of RankRC with RankSVM

In this section we analytically compare the solution of RankRC with RankSVM. In particular, we establish a bound for the difference between the solution of RankSVM and a solution in which the hypothesis is restricted to an arbitrary subset of kernel functions. This bound shows that it is optimal to first include kernel functions that correspond to points from the rare class when the dataset is unbalanced. Hence, this bound provides additional theoretical justification for RankRC.

We first establish equivalence between instantiating a hypothesis using a subset of training points and instantiating a hypothesis using the full training set but with the feature mapping equal to the orthogonal projection of the original mapping. We show this is true for an arbitrary loss function. Subsequently, we use this result to bound the difference between the RankSVM classifier and a classifier that is restricted to a subset of kernel functions, by conducting a stability analysis for the RankSVM optimization problem under a projected feature map perturbation.

For the purpose of analysis, we shall work explicitly in the high-dimensional feature space. Let $\phi : \mathcal{X} \to \mathcal{F} \subseteq \mathbb{R}^{d'}$, denote a feature map corresponding to the kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, such

that $\forall \mathbf{u}, \mathbf{v} \in \mathcal{X}$, $k(\mathbf{u}, \mathbf{v}) = \phi(\mathbf{u})^T \phi(\mathbf{v})$. For an arbitrary loss function, $L : \mathbb{R}^m \to \mathbb{R}$, and regularization parameter, $\lambda \in \mathbb{R}_+$, consider the following regularized loss minimization problem in space $\mathcal{F}$,

$$\min_{\mathbf{w} \in \mathbb{R}^{d'}} \quad L\left(\mathbf{w}^T \phi(\mathbf{x}_1), ..., \mathbf{w}^T \phi(\mathbf{x}_m)\right) + \frac{\lambda}{2} \|\mathbf{w}\|_2^2 \,. \tag{11}$$

Here, the hypothesis, $f_\phi(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x})$, can be nonlinear in the input space, $\mathcal{X}$, but is linear in the high-dimensional feature space, $\mathcal{F}$. We use the subscript ϕ in f_ϕ to indicate the feature map used in the hypothesis. Note, RankSVM is a special case of (11) using a ranking loss for L. From the Representer Theorem, the solution of (11) is of the form

$$f_\phi(\mathbf{x}) = \sum_{i=1}^{m} \beta_i k(\mathbf{x}_i, \mathbf{x}) = \sum_{i=1}^{m} \beta_i \phi(\mathbf{x}_i)^T \phi(\mathbf{x}) \,, \quad \text{or} \quad \mathbf{w} = \sum_{i=1}^{m} \beta_i \phi(\mathbf{x}_i) \,. \tag{12}$$

This implies that the optimal hypothesis can always be represented using the full training set and the solution vector $\mathbf{w} \in \mathcal{S} = \text{span}\{\phi(\mathbf{x}_i) : i = 1, ..., m\}$ is a linear combination of all the points in feature space.

Now consider restricting the hypothesis to an arbitrary subset of kernel functions, indexed by $\mathcal{R} \subseteq \{1, ..., m\}$:

$$\bar{f}_\phi(\mathbf{x}) = \sum_{i \in \mathcal{R}} \beta_i k(\mathbf{x}_i, \mathbf{x}) = \sum_{i \in \mathcal{R}} \beta_i \phi(\mathbf{x}_i)^T \phi(\mathbf{x}) \,, \quad \text{or} \quad \mathbf{w} = \sum_{i \in \mathcal{R}} \beta_i \phi(\mathbf{x}_i) \,, \tag{13}$$

with $\bar{f}_\phi(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x})$. We use the overline in $\bar{f}_\phi$ to indicate a restricted hypothesis. Subsequently, we shall refer to (13) as the $\mathcal{R}$-subset representation or classifier. In this case, the solution vector, $\mathbf{w} \in \mathcal{S}_\mathcal{R} = \text{span}\{\phi(\mathbf{x}_i) : i \in \mathcal{R}\}$, is a linear combination of the subset of points in feature space indexed by $\mathcal{R}$. Since the set $\mathcal{S}_\mathcal{R}$ defines all feasible values of $\mathbf{w}$, restricting the hypothesis to the $\mathcal{R}$-subset representation corresponds to solving the following *constrained* regularized loss minimization problem in feature space:

$$\begin{aligned} \min_{\mathbf{w} \in \mathbb{R}^{d'}} \quad & L\left(\mathbf{w}^T \phi(\mathbf{x}_1), ..., \mathbf{w}^T \phi(\mathbf{x}_m)\right) + \frac{\lambda}{2} \|\mathbf{w}\|_2^2 \,, \\ \text{subject to} \quad & \mathbf{w} \in \mathcal{S}_\mathcal{R} = \text{span}\{\phi(\mathbf{x}_i) : i \in \mathcal{R}\} \,, \quad \mathcal{R} \subseteq \{1, ..., m\} \,. \end{aligned} \tag{14}$$

Note, RankRC is a special case of (14) using a ranking loss for L and setting $\mathcal{R} = \{i : y_i = 1\}$.

In Theorem 2 we will establish that problem (14) is equivalent to problem (11) under a projected feature map. That is, problem (14) is equivalent to the following *unconstrained* loss minimization problem,

$$\min_{\mathbf{w} \in \mathbb{R}^{d'}} \quad L\left(\mathbf{w}^T \phi_\mathcal{R}(\mathbf{x}_1), ..., \mathbf{w}^T \phi_\mathcal{R}(\mathbf{x}_m)\right) + \frac{\lambda}{2} \|\mathbf{w}\|_2^2 \,, \tag{15}$$

with hypothesis, $f_{\phi_{\mathcal{R}}}(\mathbf{x}) = \mathbf{w}^T \phi_{\mathcal{R}}(\mathbf{x})$ and a feature map, $\phi_{\mathcal{R}} : \mathcal{X} \to \mathcal{F}_{\mathcal{R}} \subseteq \mathbb{R}^{d'}$, defined as the orthogonal projection of ϕ onto $\mathcal{S}_{\mathcal{R}}$, i.e.,

$$\phi_{\mathcal{R}}(\mathbf{x}) = \text{Proj}_{\mathcal{S}_{\mathcal{R}}}(\phi(\mathbf{x})) \ . \tag{16}$$

The feature map, $\phi_{\mathcal{R}}$, maps the input space to a feature space, $\mathcal{F}_{\mathcal{R}}$, which contains vectors of the same dimensionality, d', as the original feature space, $\mathcal{F}$. Before establishing the equivalence of (14) and (15), we first prove a technical lemma.

Lemma 1. *Consider a feature map, $\phi : \mathcal{X} \to \mathcal{F} \subseteq \mathbb{R}^{d'}$, and its projected map, $\phi_{\mathcal{R}} : \mathcal{X} \to \mathcal{F}_{\mathcal{R}} \subseteq \mathbb{R}^{d'}$, defined by* (16) *for some index subset $\mathcal{R} \subseteq \{1, ..., m\}$ and $\mathcal{S}_{\mathcal{R}} = span\{\phi(\mathbf{x}_i) : i \in \mathcal{R}\}$. Assume that $\mathbf{w} \in \mathbb{R}^{d'}$ is feasible for the constrained regularized loss minimization problem problem* (14)*. Let $\bar{f}_\phi(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x})$ and $f_{\phi_{\mathcal{R}}}(\mathbf{x}) = \mathbf{w}^T \phi_{\mathcal{R}}(\mathbf{x})$ be hypotheses associated with feature mapping ϕ and $\phi_{\mathcal{R}}$, respectively. Then*

$$\bar{f}_\phi(\mathbf{x}) = f_{\phi_{\mathcal{R}}}(\mathbf{x}), \quad \forall \mathbf{x} \in \mathcal{X}. \tag{17}$$

PROOF. Given any $\phi(\mathbf{x})$, there exists a unique orthogonal decomposition

$$\phi(\mathbf{x}) = \phi_{\mathcal{R}}(\mathbf{x}) + \phi^{\perp}_{\mathcal{R}}(\mathbf{x}) \ , \tag{18}$$

where $\phi_{\mathcal{R}}(\mathbf{x}) \in \mathcal{S}_{\mathcal{R}} \subseteq \mathbb{R}^{d'}$ is a component in $\mathcal{S}_{\mathcal{R}}$ and $\phi^{\perp}_{\mathcal{R}}(\mathbf{x}) \in \mathbb{R}^{d'}$ is a component orthogonal to $\mathcal{S}_{\mathcal{R}}$. By definition, $\phi(\mathbf{x}_i) \in \mathcal{S}_{\mathcal{R}}, \forall i \in \mathcal{R}$. Hence

$$\phi(\mathbf{x}_i)^T \phi^{\perp}_{\mathcal{R}}(\mathbf{x}) = 0, \quad \forall i \in \mathcal{R}, \ \forall \mathbf{x} \in \mathcal{X}. \tag{19}$$

Since $\mathbf{w}$ is a feasible point for (14), we can write $\mathbf{w} = \sum_{i \in \mathcal{R}} \beta_i \phi(\mathbf{x}_i)$ for some $\boldsymbol{\beta} \in \mathbb{R}^{|\mathcal{R}|}$. Then using (19) we have

$$\begin{aligned}
\bar{f}_\phi(\mathbf{x}) &= \mathbf{w}^T \phi(\mathbf{x}) \\
&= \left(\sum_{i \in \mathcal{R}} \beta_i \phi(\mathbf{x}_i) \right)^T \left(\phi_{\mathcal{R}}(\mathbf{x}) + \phi^{\perp}_{\mathcal{R}}(\mathbf{x}) \right) \\
&= \left(\sum_{i \in \mathcal{R}} \beta_i \phi(\mathbf{x}_i) \right)^T \phi_{\mathcal{R}}(\mathbf{x}) \\
&= \mathbf{w}^T \phi_{\mathcal{R}}(\mathbf{x}) \\
&= f_{\phi_{\mathcal{R}}}(\mathbf{x}).
\end{aligned}$$

This completes the proof. □

Lemma 1 shows that, for any feasible $\mathbf{w}$ of (14), the hypothesis $\bar{f}_\phi(\mathbf{x})$ corresponding to the map ϕ, is equivalent to the hypothesis $f_{\phi_\mathcal{R}}(\mathbf{x})$, corresponding to the projected map $\phi_\mathcal{R}$.

Theorem 2. *Consider a feature map, $\phi : \mathcal{X} \to \mathcal{F} \subseteq \mathbb{R}^{d'}$, and its projected map, $\phi_\mathcal{R} : \mathcal{X} \to \mathcal{F}_\mathcal{R} \subseteq \mathbb{R}^{d'}$, defined by* (16) *for some index subset $\mathcal{R} \subseteq \{1, ..., m\}$ and $\mathcal{S}_\mathcal{R} = span\{\phi(\mathbf{x}_i) : i \in \mathcal{R}\}$. Then the constrained regularized loss minimization problem* (14)*, using map ϕ, is equivalent to the unconstrained regularized loss minimization problem* (15)*, using the projected map $\phi_\mathcal{R}$, i.e., $\mathbf{w}^*$ solves* (14) *if and only if $\mathbf{w}^*$ solves* (15)*. In addition, assuming $\mathbf{w}^*$ solves either* (14) *or* (15)*, then the hypothesis $\bar{f}^*_\phi(\mathbf{x}) = (\mathbf{w}^*)^T\phi(\mathbf{x})$, with map ϕ, is equivalent to the hypothesis $f^*_{\phi_\mathcal{R}}(\mathbf{x}) = (\mathbf{w}^*)^T\phi_\mathcal{R}(\mathbf{x})$, with the projected map $\phi_\mathcal{R}$, i.e.,*

$$\bar{f}^*_\phi(\mathbf{x}) = f^*_{\phi_\mathcal{R}}(\mathbf{x}), \quad \forall \mathbf{x} \in \mathcal{X}. \tag{20}$$

PROOF. Using the Representor Theorem, there exists a solution $\mathbf{w}^*_\mathcal{R}$ to problem (15), which can be expressed as

$$\mathbf{w}^*_\mathcal{R} = \sum_{i=1}^{m} \beta^*_i \phi_\mathcal{R}(\mathbf{x}_i). \tag{21}$$

Hence, for any $\mathbf{w} \in \mathbb{R}^{d'}$,

$$L\left((\mathbf{w}^*_\mathcal{R})^T\phi_\mathcal{R}(\mathbf{x}_1), ..., (\mathbf{w}^*_\mathcal{R})^T\phi_\mathcal{R}(\mathbf{x}_m)\right) + \frac{\lambda}{2}\|\mathbf{w}^*_\mathcal{R}\|_2^2 \leq L\left(\mathbf{w}^T\phi_\mathcal{R}(\mathbf{x}_1), ..., \mathbf{w}^T\phi_\mathcal{R}(\mathbf{x}_m)\right) + \frac{\lambda}{2}\|\mathbf{w}\|_2^2 \tag{22}$$

Since $\phi_\mathcal{R}(\mathbf{x}) \in \mathcal{S}_\mathcal{R}$, from (21), $\mathbf{w}^*_\mathcal{R} \in \mathcal{S}_\mathcal{R}$. Hence $\mathbf{w}^*_\mathcal{R}$ satisfies the constraint in (14). Following Lemma 1,

$$(\mathbf{w}^*_\mathcal{R})^T\phi_\mathcal{R}(\mathbf{x}) = (\mathbf{w}^*_\mathcal{R})^T\phi(\mathbf{x}), \quad \forall \mathbf{x} \in \mathcal{X}. \tag{23}$$

Now consider any feasible point $\mathbf{w}$ for (14). Following Lemma 1, we have

$$\mathbf{w}^T\phi(\mathbf{x}) = \mathbf{w}^T\phi_\mathcal{R}(\mathbf{x}), \quad \forall \mathbf{x} \in \mathcal{X}. \tag{24}$$

From (22) and (24),

$$L\left((\mathbf{w}^*_\mathcal{R})^T\phi_\mathcal{R}(\mathbf{x}_1), ..., (\mathbf{w}^*_\mathcal{R})^T\phi_\mathcal{R}(\mathbf{x}_m)\right) + \frac{\lambda}{2}\|\mathbf{w}^*_\mathcal{R}\|_2^2 \leq L\left(\mathbf{w}^T\phi(\mathbf{x}_1), ..., \mathbf{w}^T\phi(\mathbf{x}_m)\right) + \frac{\lambda}{2}\|\mathbf{w}\|_2^2 \tag{25}$$

Hence $\mathbf{w}^*_\mathcal{R}$ is a solution to (14).

Conversely let us assume that $\mathbf{w}^*$ is a solution to (14). Since $\mathbf{w}^*_\mathcal{R}$ is feasible for (14),

$$
\begin{aligned}
L\left((\mathbf{w}^*)^T\phi(\mathbf{x}_1),...,(\mathbf{w}^*)^T\phi(\mathbf{x}_m)\right)+\frac{\lambda}{2}\|\mathbf{w}^*\|_2^2 &\leq L\left((\mathbf{w}^*_{\mathcal{R}})^T\phi(\mathbf{x}_1),...,(\mathbf{w}^*_{\mathcal{R}})^T\phi(\mathbf{x}_m)\right)+\frac{\lambda}{2}\|\mathbf{w}^*_{\mathcal{R}}\|_2^2 \\
&= L\left((\mathbf{w}^*_{\mathcal{R}})^T\phi_{\mathcal{R}}(\mathbf{x}_1),...,(\mathbf{w}^*_{\mathcal{R}})^T\phi_{\mathcal{R}}(\mathbf{x}_m)\right)+\frac{\lambda}{2}\|\mathbf{w}^*_{\mathcal{R}}\|_2^2.
\end{aligned}
$$

where the equality follows from (23).

From Lemma 1,

$$
(\mathbf{w}^*)^T\phi(\mathbf{x})=(\mathbf{w}^*)^T\phi_{\mathcal{R}}(\mathbf{x}),\quad \forall\mathbf{x}\in\mathcal{X}.
$$

Hence

$$
L\left((\mathbf{w}^*)^T\phi_{\mathcal{R}}(\mathbf{x}_1),...,(\mathbf{w}^*)^T\phi_{\mathcal{R}}(\mathbf{x}_m)\right)+\frac{\lambda}{2}\|\mathbf{w}^*\|_2^2\leq L\left((\mathbf{w}^*_{\mathcal{R}})^T\phi_{\mathcal{R}}(\mathbf{x}_1),...,(\mathbf{w}^*_{\mathcal{R}})^T\phi_{\mathcal{R}}(\mathbf{x}_m)\right)+\frac{\lambda}{2}\|\mathbf{w}^*_{\mathcal{R}}\|_2^2.
$$

Following (22) , for any $\mathbf{w}\in\mathbb{R}^{d'}$,

$$
L\left((\mathbf{w}^*)^T\phi_{\mathcal{R}}(\mathbf{x}_1),...,(\mathbf{w}^*)^T\phi_{\mathcal{R}}(\mathbf{x}_m)\right)+\lambda\|\mathbf{w}^*\|_2^2\leq L\left(\mathbf{w}^T_{\mathcal{R}}\phi_{\mathcal{R}}(\mathbf{x}_1),...,\mathbf{w}^T_{\mathcal{R}}\phi_{\mathcal{R}}(\mathbf{x}_m)\right)+\frac{\lambda}{2}\|\mathbf{w}\|_2^2
$$

Hence the solution $\mathbf{w}^*$ is also a solution to (15). The result (20) immediately follows from Lemma 1 and the equivalence of (14) and (15). The proof is complete. □

Now consider the ranking loss problem. Define

$$
R_\phi(\mathbf{w})=\frac{1}{m_+m_-}\sum_{\{i:y_i=+1\}}\sum_{\{j:y_j=-1\}}\ell_h\left(\mathbf{w}^T\phi(\mathbf{x}_i)-\mathbf{w}^T\phi(\mathbf{x}_j)\right), \tag{26}
$$

where $\ell_h(z)=\max(0,1-z)$ is the hinge loss. Setting

$$
L\left(\mathbf{w}^T\phi(\mathbf{x}_1),...,\mathbf{w}^T\phi(\mathbf{x}_m)\right)=R_\phi(\mathbf{w}) \tag{27}
$$

in (11) gives:

$$
\min_{\mathbf{w}\in\mathbb{R}^{d'}}\quad F_\phi(\mathbf{w})=R_\phi(\mathbf{w})+\frac{\lambda}{2}\|\mathbf{w}\|_2^2\,. \tag{28}
$$

Problem (28) corresponds to the RankSVM problem in feature space $\mathcal{F}$, defined by the feature map ϕ (or implicitly, by the kernel function, k). Similarly, setting (27) in problem (14) corresponds to a RankSVM problem in which the hypothesis is restricted to a $\mathcal{R}$-subset representation:

$$
\begin{aligned}
\min_{\mathbf{w}\in\mathbb{R}^{d'}}\quad & F_\phi(\mathbf{w})=R_\phi(\mathbf{w})+\frac{\lambda}{2}\|\mathbf{w}\|_2^2\,, \\
\text{subject to}\quad & \mathbf{w}\in\mathcal{S}_{\mathcal{R}}=\text{span}\{\phi(\mathbf{x}_i):i\in\mathcal{R}\}\,,\quad \mathcal{R}\subseteq\{1,...,m\}\,.
\end{aligned} \tag{29}
$$

Setting $L\left(\mathbf{w}^T\phi_{\mathcal{R}}(\mathbf{x}_1),...,\mathbf{w}^T\phi_{\mathcal{R}}(\mathbf{x}_m)\right) = R_{\phi_{\mathcal{R}}}(\mathbf{w})$ in problem (15), corresponds to the RankSVM problem with feature map $\phi_{\mathcal{R}}$:

$$\min_{\mathbf{w}\in\mathbb{R}^{d'}} \quad F_{\phi_{\mathcal{R}}}(\mathbf{w}) = R_{\phi_{\mathcal{R}}}(\mathbf{w}) + \frac{\lambda}{2}\|\mathbf{w}\|_2^2 \,. \tag{30}$$

Let f^*_ϕ, $\bar{f}^*_\phi$ and $f^*_{\phi_{\mathcal{R}}}$ denote the optimal hypotheses obtained by solving (28), (29) and (30), respectively. From Theorem 2, $\bar{f}^*_\phi(\mathbf{x}) = f^*_{\phi_{\mathcal{R}}}(\mathbf{x})$, and therefore $|f^*_\phi(\mathbf{x}) - \bar{f}^*_\phi(\mathbf{x})| = |f^*_\phi(\mathbf{x}) - f^*_{\phi_{\mathcal{R}}}(\mathbf{x})|$. In other words, we can bound the difference between a RankSVM classifier and a $\mathcal{R}$-subset classifier, by a stability analysis of the optimal RankSVM hypothesis under a perturbed (projected) feature map.

Stability analyses for a regular SVM have been conducted previously. In particular, Bousquet and Elisseeff (2002) obtain a bound for a regular SVM under the effect of changing one training point. Cortes et al. (2010) analyze stability of a regular SVM under the effect of changing the kernel matrix. Our stability analysis here differs from existing analyses in two aspects. Firstly, we obtain a bound under the effect of changing the feature mapping ϕ to $\phi_{\mathcal{R}}$. Secondly, we consider here the RankSVM problem instead of a regular SVM.

Theorem 3. *Consider a feature map, $\phi : \mathcal{X} \to \mathcal{F} \subseteq \mathbb{R}^{d'}$, and its projected map, $\phi_{\mathcal{R}} : \mathcal{X} \to \mathcal{F}_{\mathcal{R}} \subseteq \mathbb{R}^{d'}$, defined by* (16) *for some index subset $\mathcal{R} \subseteq \{1,...,m\}$ and $\mathcal{S}_{\mathcal{R}} = span\{\phi(\mathbf{x}_i) : i \in \mathcal{R}\}$. Assume that $f^*_\phi(\mathbf{x}) = (\mathbf{w}^*)^T\phi(\mathbf{x})$ is the optimal RankSVM hypothesis obtained by solving* (28) *with feature map ϕ, and $f^*_{\phi_{\mathcal{R}}}(\mathbf{x}) = (\mathbf{w}^*_{\mathcal{R}})^T\phi_{\mathcal{R}}(\mathbf{x})$ is the optimal RankSVM hypothesis obtained by solving* (30) *with feature map $\phi_{\mathcal{R}}$. Assume there exists $\kappa > 0$ such that $k(\mathbf{x},\mathbf{x}) \le \kappa$, where $k : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ is the kernel map associated with ϕ. Then the following inequality holds,*

$$|f^*_{\phi_{\mathcal{R}}}(\mathbf{x}) - f^*_\phi(\mathbf{x})| \le \frac{2\kappa}{\lambda}\left(\sum_{\{i:y_i=+1\}}\frac{\mathbb{I}[i\notin\mathcal{R}]}{m_+} + \sum_{\{j:y_j=-1\}}\frac{\mathbb{I}[j\notin\mathcal{R}]}{m_-}\right)^{\frac{1}{2}}, \quad \forall\mathbf{x}\in\mathcal{X}, \tag{31}$$

where $\mathbb{I}[p]$ denotes the indicator function and is equal to 1 if p is true, 0 if p is false.

PROOF. Assume that $\mathbf{w}^*$ and $\mathbf{w}^*_{\mathcal{R}}$ are minimizers of (28) and (30), respectively. Let $\Delta\mathbf{w} = \mathbf{w}^*_{\mathcal{R}} - \mathbf{w}^*$.

Recall that a convex function g satisfies

$$g(\mathbf{u} + t(\mathbf{v}-\mathbf{u})) - g(\mathbf{u}) \le t(g(\mathbf{v}) - g(\mathbf{u}))$$

for all $\mathbf{u},\mathbf{v},t \in [0,1]$. Since ℓ_h is convex, R_ϕ and $R_{\phi_{\mathcal{R}}}$ are convex. Then

$$R_\phi(\mathbf{w}^* + t\Delta\mathbf{w}) - R_\phi(\mathbf{w}^*) \le t(R_\phi(\mathbf{w}^*_{\mathcal{R}}) - R_\phi(\mathbf{w}^*)) \tag{32}$$

$$\text{and} \quad R_{\phi_{\mathcal{R}}}(\mathbf{w}^*_{\mathcal{R}} - t\Delta\mathbf{w}) - R_{\phi_{\mathcal{R}}}(\mathbf{w}^*_{\mathcal{R}}) \le t(R_{\phi_{\mathcal{R}}}(\mathbf{w}^*) - R_{\phi_{\mathcal{R}}}(\mathbf{w}^*_{\mathcal{R}})) \,, \tag{33}$$

for all $t \in [0,1]$.

Since $\mathbf{w}^*$ and $\mathbf{w}^*_{\mathcal{R}}$ are minimizers of F_ϕ and $F_{\phi_{\mathcal{R}}}$, for any $t \in [0,1]$, we have

$$F_\phi(\mathbf{w}^*) \le F_\phi(\mathbf{w}^* + t\Delta\mathbf{w}) \tag{34}$$

$$\text{and} \quad F_{\phi_{\mathcal{R}}}(\mathbf{w}^*_{\mathcal{R}}) \le F_{\phi_{\mathcal{R}}}(\mathbf{w}^*_{\mathcal{R}} - t\Delta\mathbf{w})\,. \tag{35}$$

Summing (34) and (35), using $F_\phi(\mathbf{w}) = R_\phi(\mathbf{w}) + \frac{\lambda}{2}\|\mathbf{w}\|_2^2$ and the identity

$$\left(\|\mathbf{w}^*\|^2 - \|\mathbf{w}^* + t\Delta\mathbf{w}\|^2\right) + \left(\|\mathbf{w}^*_{\mathcal{R}}\|^2 - \|\mathbf{w}^*_{\mathcal{R}} - t\Delta\mathbf{w}\|^2\right) = 2t(1-t)\|\Delta\mathbf{w}\|^2,$$

we obtain

$$\lambda t(1-t)\|\Delta\mathbf{w}\|^2 \le \left(R_\phi(\mathbf{w}^* + t\Delta\mathbf{w}) - R_\phi(\mathbf{w}^*)\right) + \left(R_{\phi_{\mathcal{R}}}(\mathbf{w}^*_{\mathcal{R}} - t\Delta\mathbf{w}) - R_{\phi_{\mathcal{R}}}(\mathbf{w}^*_{\mathcal{R}})\right) \tag{36}$$

Substituting (32) and (33) into (36), dividing by λt, and taking the limit $t \to 0$ gives

$$\begin{aligned}
\|\Delta\mathbf{w}\|^2 &\le \frac{1}{\lambda}\left(R_\phi(\mathbf{w}^*_{\mathcal{R}}) - R_{\phi_{\mathcal{R}}}(\mathbf{w}^*_{\mathcal{R}}) + R_{\phi_{\mathcal{R}}}(\mathbf{w}^*) - R_\phi(\mathbf{w}^*)\right) \\
&= \frac{1}{\lambda m_+ m_-} \sum_{\{i:y_i=+1\}} \sum_{\{j:y_j=-1\}} \Big[\ell_h\left((\mathbf{w}^*_{\mathcal{R}})^T\phi(\mathbf{x}_i) - (\mathbf{w}^*_{\mathcal{R}})^T\phi(\mathbf{x}_j)\right) - \ell_h\left((\mathbf{w}^*_{\mathcal{R}})^T\phi_{\mathcal{R}}(\mathbf{x}_i) - (\mathbf{w}^*_{\mathcal{R}})^T\phi_{\mathcal{R}}(\mathbf{x}_j)\right) \\
&\qquad + \ell_h\left((\mathbf{w}^*)^T\phi_{\mathcal{R}}(\mathbf{x}_i) - (\mathbf{w}^*)^T\phi_{\mathcal{R}}(\mathbf{x}_j)\right) - \ell_h\left((\mathbf{w}^*)^T\phi(\mathbf{x}_i) - (\mathbf{w}^*)^T\phi(\mathbf{x}_j)\right)\Big]\,,
\end{aligned}$$

where the last inequality uses the definitions of R_ϕ and $R_{\phi_{\mathcal{R}}}$ respectively. Since $\ell_h(\cdot)$ is 1-Lipschitz, we obtain

$$\begin{aligned}
\|\Delta\mathbf{w}\|^2 &\le \frac{1}{\lambda m_+ m_-} \sum_{\{i:y_i=+1\}} \sum_{\{j:y_j=-1\}} (\|\mathbf{w}^*\| + \|\mathbf{w}^*_{\mathcal{R}}\|)\Big(\|\phi(\mathbf{x}_i) - \phi_{\mathcal{R}}(\mathbf{x}_i)\| + \|\phi(\mathbf{x}_j) - \phi_{\mathcal{R}}(\mathbf{x}_j)\|\Big) \\
&= \frac{\|\mathbf{w}^*\| + \|\mathbf{w}^*_{\mathcal{R}}\|}{\lambda}\left(\sum_{\{i:y_i=+1\}} \frac{\|\phi(\mathbf{x}_i) - \phi_{\mathcal{R}}(\mathbf{x}_i)\|}{m_+} + \sum_{\{j:y_j=-1\}} \frac{\|\phi(\mathbf{x}_j) - \phi_{\mathcal{R}}(\mathbf{x}_j)\|}{m_-}\right)\,.
\end{aligned} \tag{37}$$

From $\phi(\mathbf{x}) = \phi_{\mathcal{R}}(\mathbf{x}) + \phi^\perp_{\mathcal{R}}(\mathbf{x})$, with $\phi_{\mathcal{R}}(\mathbf{x}) \in \mathcal{S}_{\mathcal{R}}$ and $\phi^\perp_{\mathcal{R}}(\mathbf{x})$ is in the space orthogonal to $\mathcal{S}_{\mathcal{R}}$, we have, for $i = 1, \ldots, m$,

$$\|\phi(\mathbf{x}_i) - \phi_{\mathcal{R}}(\mathbf{x}_i)\| = \|\phi^\perp_{\mathcal{R}}(\mathbf{x}_i)\| \le \begin{cases} \|\phi(\mathbf{x}_i)\| = \sqrt{k(\mathbf{x}_i, \mathbf{x}_i)} \le \sqrt{\kappa}, & \text{if } i \notin \mathcal{R} \\ 0, & \text{if } i \in \mathcal{R}\,. \end{cases} \tag{38}$$

In addition, recall that RankSVM is equivalent to a 1-class SVM on an enlarged dataset with the set of points $\mathcal{P} = \{\phi(\mathbf{x}_i) - \phi(\mathbf{x}_j) : y_i > y_j, i, j = 1..., m\}$. Therefore $\mathbf{w}$ can be expressed in terms of the dual variables $0 \leq \alpha^*_{ij} \leq C$ of an SVM problem trained on $\mathcal{P}$ with $C = \frac{1}{\lambda m_+ m_-}$, as follows,

$$\begin{aligned}\mathbf{w}^* &= \sum_{\{i,j:y_i>y_j\}} \alpha^*_{ij}(\phi(\mathbf{x}_i) - \phi(\mathbf{x}_j)) = \sum_{\{i:y_i=+1\}} \sum_{\{j:y_j=-1\}} \alpha^*_{ij}(\phi(\mathbf{x}_i) - \phi(\mathbf{x}_j)) \\ &= \sum_{\{i:y_i=+1\}} \phi(\mathbf{x}_i) \left(\sum_{\{j:y_j=-1\}} \alpha^*_{ij} \right) - \sum_{\{j:y_j=-1\}} \phi(\mathbf{x}_j) \left(\sum_{\{i:y_i=+1\}} \alpha^*_{ij} \right) .\end{aligned}$$

Since $\|\phi(\mathbf{x})\| \leq \sqrt{\kappa}$ and $C = \frac{1}{\lambda m_+ m_-}$, we get $\|\mathbf{w}^*\| \leq \sqrt{\kappa} C m_- m_+ + \sqrt{\kappa} C m_+ m_- = \frac{2\sqrt{\kappa}}{\lambda}$. Similarly, $\|\phi_{\mathcal{R}}(\mathbf{x})\| \leq \|\phi(\mathbf{x})\| \leq \sqrt{\kappa}$ and $\|\mathbf{w}^*_{\mathcal{R}}\| \leq \frac{2\sqrt{\kappa}}{\lambda}$. Together with (38), we can then bound (37) by

$$\|\Delta \mathbf{w}\|^2 \leq \frac{4\kappa}{\lambda^2} \left(\sum_{\{i:y_i=+1\}} \frac{\mathbb{I}[i \notin \mathcal{R}]}{m_+} + \sum_{\{j:y_j=-1\}} \frac{\mathbb{I}[j \notin \mathcal{R}]}{m_-} \right) .$$

Therefore, we obtain

$$\begin{aligned}|f_{\phi_{\mathcal{R}}}(\mathbf{x}) - f_\phi(\mathbf{x})| &= |\mathbf{w}^T_{\mathcal{R}} \phi_{\mathcal{R}}(\mathbf{x}) - \mathbf{w}^T \phi(\mathbf{x})| \\ &= |\mathbf{w}^T_{\mathcal{R}} \left(\phi(\mathbf{x}) - \phi^{\perp}_{\mathcal{R}}(\mathbf{x}) \right) - \mathbf{w}^T \phi(\mathbf{x})| \\ &= |\Delta \mathbf{w}^T \phi(\mathbf{x}) - \mathbf{w}^T_{\mathcal{R}} \phi^{\perp}_{\mathcal{R}}(\mathbf{x})| \\ &= |\Delta \mathbf{w}^T \phi(\mathbf{x})| \\ &\leq \|\Delta \mathbf{w}\| \|\phi(\mathbf{x})\| \\ &\leq \frac{2\kappa}{\lambda} \left(\sum_{\{i:y_i=+1\}} \frac{\mathbb{I}[i \notin \mathcal{R}]}{m_+} + \sum_{\{j:y_j=-1\}} \frac{\mathbb{I}[j \notin \mathcal{R}]}{m_-} \right)^{\frac{1}{2}} ,\end{aligned}$$

where we have used $\mathbf{w}^T_{\mathcal{R}} \phi^{\perp}_{\mathcal{R}}(\mathbf{x}) = 0$ in the third equality since $\mathbf{w}_{\mathcal{R}} \in \mathcal{S}_{\mathcal{R}}$. This completes the proof. □

The following result is a direct consequence of Theorem 2 and Theorem 3.

Corollary 4. *For a feature map, $\phi : \mathcal{X} \to \mathcal{F} \subseteq \mathbb{R}^{d'}$ associated with kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, let $f^*_\phi(\mathbf{x})$ be the optimal RankSVM hypothesis obtained by solving* (28)*, and $\bar{f}^*_\phi(\mathbf{x})$ be the optimal hypothesis obtained by solving* (29)*, in which the hypothesis is restricted to an arbitrary subset of kernel*

functions indexed by $\mathcal{R} \subseteq \{1,...,m\}$. *Assume there exists* $\kappa > 0$ *such that* $k(\mathbf{x},\mathbf{x}) \leq \kappa$. *Then the following inequality holds,*

$$|\bar{f}^*_\phi(\mathbf{x}) - f^*_\phi(\mathbf{x})| \leq \frac{2\kappa}{\lambda}\left(\sum_{\{i:y_i=+1\}} \frac{\mathbb{I}[i \notin \mathcal{R}]}{m_+} + \sum_{\{j:y_j=-1\}} \frac{\mathbb{I}[j \notin \mathcal{R}]}{m_-}\right)^{\frac{1}{2}}, \quad \forall \mathbf{x} \in \mathcal{X}. \tag{39}$$

□

Therefore, for the ranking loss, the bound (39) decreases asymmetrically depending on whether we include a point from the positive or negative class. In particular, if the dataset is unbalanced with $m_- \gg m_+$, or $\frac{1}{m_+} \gg \frac{1}{m_-}$, then the reduction obtained from including a positive class kernel function is much greater than including one from the negative class. Hence, for a fixed number of kernel functions, the bound is minimized by first including kernel functions corresponding to the positive or rare class.

4. Relation to Nyström Approximation

The Nyström method approximates a symmetric positive semi-definite matrix $Q \in \mathbb{R}^{m\times m}$ by a sample submatrix D of $n \ll m$ columns from Q (e.g. see Baker, 1977; Williams and Seeger, 2001). Without loss of generality, assume that the first n columns are the randomly chosen samples. Then D and Q can be written as

$$D = \begin{bmatrix} A \\ B \end{bmatrix} \quad \text{and} \quad Q = \begin{bmatrix} A & B^T \\ B & C \end{bmatrix},$$

with $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{(m-n)\times n}$, and $C \in \mathbb{R}^{(m-n)\times(m-n)}$. The Nyström method computes a rank-n approximation of Q as

$$\hat{Q} = DA^\dagger D^T = \begin{bmatrix} A & B^T \\ B & BA^\dagger B^T \end{bmatrix},$$

where $A^\dagger$ is the Moore-Penrose pseudoinverse of A. Thus, the Nyström method approximates C using $BA^\dagger B^T$ and can be seen as a method to complete matrix Q using information from only n columns.

Approximating a kernel matrix with a low-rank structured matrix to improve computational efficiency has been explored in the context of other kernel algorithms before. For instance, low-rank approximations have been used to speed up kernel PCA (Achlioptas et al., 2001), multi-dimensional scaling (Platt, 2005), spectral clustering (Fowlkes et al., 2004), manifold learning

(Talwalkar, 2010), Gaussian processes (Williams and Seeger, 2001), and support vector machines (Smola and Schölkopf, 2000; Fine and Scheinberg, 2002; Zhang et al., 2012).

In this section, we show that solving a regularized loss minimization with the $\mathcal{R}$-subset representation, is equivalent to solving the full unrestricted problem using a low-rank Nyström approximation of the kernel matrix. We show this is true for an arbitrary loss function.

Consider the regularized loss minimization problem (11) with a general loss function, $L : \mathbb{R}^m \to \mathbb{R}$. By substituting the solution (12) in (11), we can express the general regularized loss minimization problem (11) in terms of the kernel matrix, $K \in \mathbb{R}^{m\times m}$, and model variables, $\boldsymbol{\beta} \in \mathbb{R}^m$,

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^m} \quad L(K\boldsymbol{\beta}) + \frac{\lambda}{2}\boldsymbol{\beta}^T K\boldsymbol{\beta} \,. \tag{40}$$

Here, $K\boldsymbol{\beta} = [f_\phi(\mathbf{x}_1), ..., f_\phi(\mathbf{x}_m)]^T \in \mathbb{R}^m$, where $f_\phi(\mathbf{x}) = \sum_{i=1}^m \beta_i k(\mathbf{x}_i, \mathbf{x})$.

Similarly, substituting the $\mathcal{R}$-subset hypothesis (13) in (11), results in the following problem with model variables, $\boldsymbol{\beta} \in \mathbb{R}^{|\mathcal{R}|}$,

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^{|\mathcal{R}|}} \quad L(K_{\bullet\mathcal{R}}\boldsymbol{\beta}) + \frac{\lambda}{2}\boldsymbol{\beta}^T K_{\mathcal{R}\mathcal{R}}\boldsymbol{\beta} \,. \tag{41}$$

Here $K_{\bullet\mathcal{R}} = [k_{ij}]_{i=1...m, j\in\mathcal{R}} \in \mathbb{R}^{m\times|\mathcal{R}|}$, $k_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$, is a subset of columns from K, and $K_{\mathcal{R}\mathcal{R}} = [k_{ij}]_{i,j\in\mathcal{R}} \in \mathbb{R}^{|\mathcal{R}|\times|\mathcal{R}|}$ is the square submatrix of K indexed by $\mathcal{R}$ along columns and rows. We have $K_{\bullet\mathcal{R}}\boldsymbol{\beta} = [\bar{f}_\phi(\mathbf{x}_1), ..., \bar{f}_\phi(\mathbf{x}_m)]^T \in \mathbb{R}^m$, where the hypothesis, $\bar{f}_\phi(\mathbf{x}) = \sum_{i\in\mathcal{R}} \beta_i k(\mathbf{x}_i, \mathbf{x})$, is restricted to use a subset of kernel functions indexed by $\mathcal{R} \subseteq \{1, ..., m\}$.

In the following proposition, we show that problem (41) is equivalent to problem (40), where the kernel matrix K is replaced by a Nyström approximation, $K' \in \mathbb{R}^{m\times m}$.

Proposition 5. *For any loss function $L : \mathbb{R}^m \to \mathbb{R}$ and kernel matrix $K \in \mathbb{R}^{m\times m}$, the regularized loss minimization problem* (41)*, in which the hypothesis is restricted to a subset of kernel functions indexed by $\mathcal{R} \subseteq \{1, ..., m\}$, is equivalent to the unrestricted problem* (40) *under a perturbed kernel matrix corresponding to the Nyström approximation, $K' = K_{\bullet\mathcal{R}} K_{\mathcal{R}\mathcal{R}}^{\dagger} K_{\bullet\mathcal{R}}^T \in \mathbb{R}^{m\times m}$, where $K_{\bullet\mathcal{R}} \in \mathbb{R}^{m\times|\mathcal{R}|}$ are columns of K indexed by $\mathcal{R}$, and $K_{\mathcal{R}\mathcal{R}} \in \mathbb{R}^{|\mathcal{R}|\times|\mathcal{R}|}$ are rows of $K_{\bullet\mathcal{R}}$ indexed by $\mathcal{R}$.*

PROOF. Since $K_{\mathcal{R}\mathcal{R}}$ is positive semi-definite, using eigen-decomposition,

$$K_{\mathcal{R}\mathcal{R}} = U\Lambda U^T \,,$$

where U is an orthonormal matrix and Λ is a diagonal matrix of non-negative eigenvalues of $K_{\mathcal{R}\mathcal{R}}$.

Define $\mathbf{w} = \Lambda^{\frac{1}{2}} U^T \boldsymbol{\beta}$. Then $\boldsymbol{\beta} = U\Lambda^{\dagger\frac{1}{2}}\mathbf{w}$, and we can express (41) in terms of $\mathbf{w}$ as,

$$\min_{\mathbf{w}\in\mathbb{R}^{|\mathcal{R}|}} \quad L\left(K_{\bullet\mathcal{R}} U\Lambda^{\dagger\frac{1}{2}}\mathbf{w}\right) + \frac{\lambda}{2}\|\mathbf{w}\|_2^2 \,. \tag{42}$$

We recognize (42) as a problem in linear space with data points given by the rows of $K_{\bullet\mathcal{R}} U\Lambda^{*\frac{1}{2}} \in \mathbb{R}^{m\times|\mathcal{R}|}$.

Denote $[\phi'(\mathbf{x}_1), ..., \phi'(\mathbf{x}_m)]^T = K_{\bullet\mathcal{R}} U\Lambda^{\dagger\frac{1}{2}}$. Applying the Representer Theorem, the solution $f_{\phi'}(\mathbf{x}) = \mathbf{w}^T\phi'(\mathbf{x})$ can be expressed in the form,

$$f_{\phi'}(\mathbf{x}) = \left(\sum_{i=1}^{m} \beta_i \phi'(\mathbf{x}_i)\right)^T \phi'(\mathbf{x}), \;\text{ or }\; \mathbf{w} = \sum_{i=1}^{m} \beta_i \phi'(\mathbf{x}_i) \,, \tag{43}$$

Substituting (43) in problem (42) yields an equivalent problem,

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^m} \quad L\left(K'\boldsymbol{\beta}\right) + \frac{\lambda}{2}\boldsymbol{\beta}^T K'\boldsymbol{\beta} \,,$$

where $K' = \left(K_{\bullet\mathcal{R}} U\Lambda^{\dagger\frac{1}{2}}\right)\left(K_{\bullet\mathcal{R}} U\Lambda^{\dagger\frac{1}{2}}\right)^T = K_{\bullet\mathcal{R}} K_{\mathcal{R}\mathcal{R}}^{\dagger} K_{\bullet\mathcal{R}}^T \in \mathbb{R}^{m\times m}$, which we recognize as the Nyström approximation of K using the columns indexed by $\mathcal{R}$ as samples. The proof is completed. □

Proposition 5 formalizes the connection between selecting a set of points for the hypothesis representation, and using a low-rank Nyström approximation kernel for any regularized loss minimization problem which can be written in form (40). Conversely, it also shows that using a low-rank Nyström approximation kernel matrix can be viewed as selecting a $\mathcal{R}$-subset representable optimal hypothesis for problem (40). Thus, for problems which use a Nyström approximation kernel, Proposition 5 provides an efficient optimization formulation in the form (41), or in the linear space form (42), reducing problem dimension from m variables to $|\mathcal{R}|$ and space from $O(m^2)$ to $O(m|\mathcal{R}|)$.

Theoretically, Theorem 2 and Proposition 5 together imply that using a Nyström kernel approximation is equivalent to projecting the feature map ϕ onto the subspace spanned by the subset of samples in feature space. This relationship can potentially be used to analyze Nyström approximation algorithms based on a feature map projection, as in Theorem 2 & 3. For instance, Theorem 6 illustrates a stability bound that can be obtained for a regular SVM trained with a Nyström kernel approximation.

Theorem 6. *Let f_K^* denote the optimal hypothesis obtained by solving the SVM problem with a kernel matrix, $K = [k(\mathbf{x}_i, \mathbf{x}_j)]_{i,j=1}^m \in \mathbb{R}^{m\times m}$,*

$$\min_{\beta\in\mathbb{R}^m} \quad \frac{1}{m}\sum_{i=1}^m \ell_h(y_i K_{i\bullet}\beta) + \frac{\lambda}{2}\beta^T K\beta \,, \tag{44}$$

where $K_{i\bullet} \in \mathbb{R}^{1\times m}$ is the i^{th} row of K. Define a perturbed kernel matrix, $K' \in \mathbb{R}^{m\times m}$, as the Nyström approximation of K using the columns of K indexed by $\mathcal{R} \subseteq \{1,...,m\}$. Let $f_{K'}^*$ denote the optimal hypothesis obtained by solving the SVM problem (44) *with the Nyström approximation K'. Then the following inequality holds,*

$$|f_{K'}^*(\mathbf{x}) - f_K^*(\mathbf{x})| \le \frac{\sqrt{2}\kappa}{\lambda}\sqrt{1-\frac{|\mathcal{R}|}{m}} \,, \quad \forall \mathbf{x} \in \mathcal{X}.$$

PROOF. Following Proposition 5, using the Nyström approximation, K', in the SVM problem (44), is equivalent to

$$\min_{\beta\in\mathbb{R}^{\mathcal{R}}} \quad \frac{1}{m}\sum_{i=1}^m \ell_h(y_i K_{i\mathcal{R}}\beta) + \frac{\lambda}{2}\beta^T K_{\mathcal{R}\mathcal{R}}\beta \,, \tag{45}$$

where $K_{i\mathcal{R}}$ is the ith row of $K_{\bullet\mathcal{R}}$. Problem (45) solves SVM with a hypothesis restricted to a subset of kernel functions. Let ϕ denote the feature map corresponding to k. Then from Theorem 2, the optimal hypothesis $f_{K'}^*(\mathbf{x})$ equals the optimal SVM hypothesis under the projected mapping $\phi_{\mathcal{R}} = \mathrm{Proj}_{\mathcal{S}_{\mathcal{R}}}(\phi)$. Consequently, the difference between the optimal hypotheses, f_K^* and $f_{K'}^*$, can be bounded following the proof of Theorem 3; the only difference is that, instead of the ranking loss function, R_ϕ, we have the SVM loss, $\frac{1}{m}\sum_{i=1}^m \ell_h\left(y_i \mathbf{w}^T\phi(\mathbf{x}_i)\right)$. This means that, instead of (37), we have

$$\|\Delta\mathbf{w}\|^2 \le \frac{\|\mathbf{w}^*\| + \|\mathbf{w}_{\mathcal{R}}^*\|}{\lambda}\sum_{i=1}^m \frac{\|\phi(\mathbf{x}_i) - \phi_{\mathcal{R}}(\mathbf{x}_i)\|}{m} \,.$$

Similarly, it can be shown that $\|\mathbf{w}^*\| \le \frac{\sqrt{\kappa}}{\lambda}$, $\|\mathbf{w}_{\mathcal{R}}^*\| \le \frac{\sqrt{\kappa}}{\lambda}$ and consequently

$$|f_{K'}^*(\mathbf{x}) - f_K^*(\mathbf{x})| \le \frac{\sqrt{2}\kappa}{\lambda}\left(\frac{\sum_{i=1}^m \mathbb{I}[i\notin\mathcal{R}]}{m}\right)^{\frac{1}{2}} \,, \quad \forall \mathbf{x} \in \mathcal{X}.$$

□

Cortes et al. (2010) obtain a stability bound for SVM assuming an arbitrary kernel matrix perturbation. They use the bound to analyze Nyström kernel approximations. The bound obtained in Cortes et al. (2010) is a function of the spectral norm of the difference between the two kernel matrices, ie. $||K' - K||_2$. In comparison, the bound obtained in Theorem 6, based on the feature map projection for a Nyström approximation, is much simpler: it is proportional to the square root of the percentage of the points not in the hypothesis representation. In addition, since the Nyström approximation, K', is computed using the pseudo inverse of kernel submatrix, $K_{\mathcal{RR}}$, it can become arbitrarily far away from K, depending on the condition number of $K_{\mathcal{RR}}$. In contrast, the projected map approach offers a more stable, and often tighter, bound.

To demonstrate this, we compare bounds for the ranking loss problem obtained using the feature map projection and kernel matrix perturbation approaches. Below we state the stability bound for RankSVM under an arbitrary kernel perturbation following the approach in Cortes et al. (2010).

Theorem 7. *Let f_K^* and $f_{K'}^*$ denote the optimal hypothesis obtained by RankSVM when using the kernel matrix $K \in \mathbb{R}^{m\times m}$ and $K' \in \mathbb{R}^{m\times m}$, respectively. Then the following inequality holds for all $\mathbf{x} \in \mathcal{X}$:*

$$|f_{K'}^*(\mathbf{x}) - f_K^*(\mathbf{x})| \leq \frac{2\sqrt{2}\kappa^{\frac{3}{4}}}{\lambda} \|K' - K\|_2^{\frac{1}{2}} \left[1 + \left(\frac{\|K' - K\|_2}{4\kappa}\right)^{\frac{1}{4}}\right] . \tag{46}$$

The proof is essentially the same as that in Cortes et al. (2010). The idea is to use an explicit $(m+1)$-dimension feature map ϕ and ϕ' associated with K and K' defined according to

$$\phi(\mathbf{x}_i) = K_{m+1}^{\frac{1}{2}} \mathbf{e}_i \quad \text{and} \quad \phi'(\mathbf{x}_i) = K_{m+1}'^{\frac{1}{2}} \mathbf{e}_i \ ,$$

where K_{m+1} and K'_{m+1} are augmented versions of K and K' with the $(m+1)$th point representing an arbitrary test point, and $\mathbf{e}_i \in \mathbb{R}^{m+1}$ is a unit vector, with the ith component equal to 1, and 0 everywhere else. Then using the fact the solution is at a minimizer and the objective is convex (as done in Theorem 3), $\|\Delta \mathbf{w}\|^2$ can be bounded in terms of $\|\phi(\mathbf{x}_i) - \phi'(\mathbf{x}_i)\|_2 \leq \|K_{m+1}^{\frac{1}{2}} - K_{m+1}'^{\frac{1}{2}}\|_2 \leq \|K_{m+1} - K'_{m+1}\|_2^{\frac{1}{2}} = \|K - K'\|_2^{\frac{1}{2}}$, which can then be used to obtain the final result. The bound obtained for RankSVM under a perturbed kernel matrix is simply twice that obtained in Cortes et al. (2010) for a regular SVM. The factor of two emerges due to the double summation in the ranking loss function.

From Proposition 5, we can bound the difference between the RankSVM classifier and a $\mathcal{R}$-subset RankSVM classifier by comparing the effect of perturbing the kernel matrix to its Nyström

approximation. Thus, bound (46) also applies to the difference between the RankSVM classifier and a $\mathcal{R}$-subset RankSVM classifier by setting $K' = K_{\bullet\mathcal{R}} K_{\mathcal{R}\mathcal{R}}^{\dagger} K_{\bullet\mathcal{R}}^{T}$.

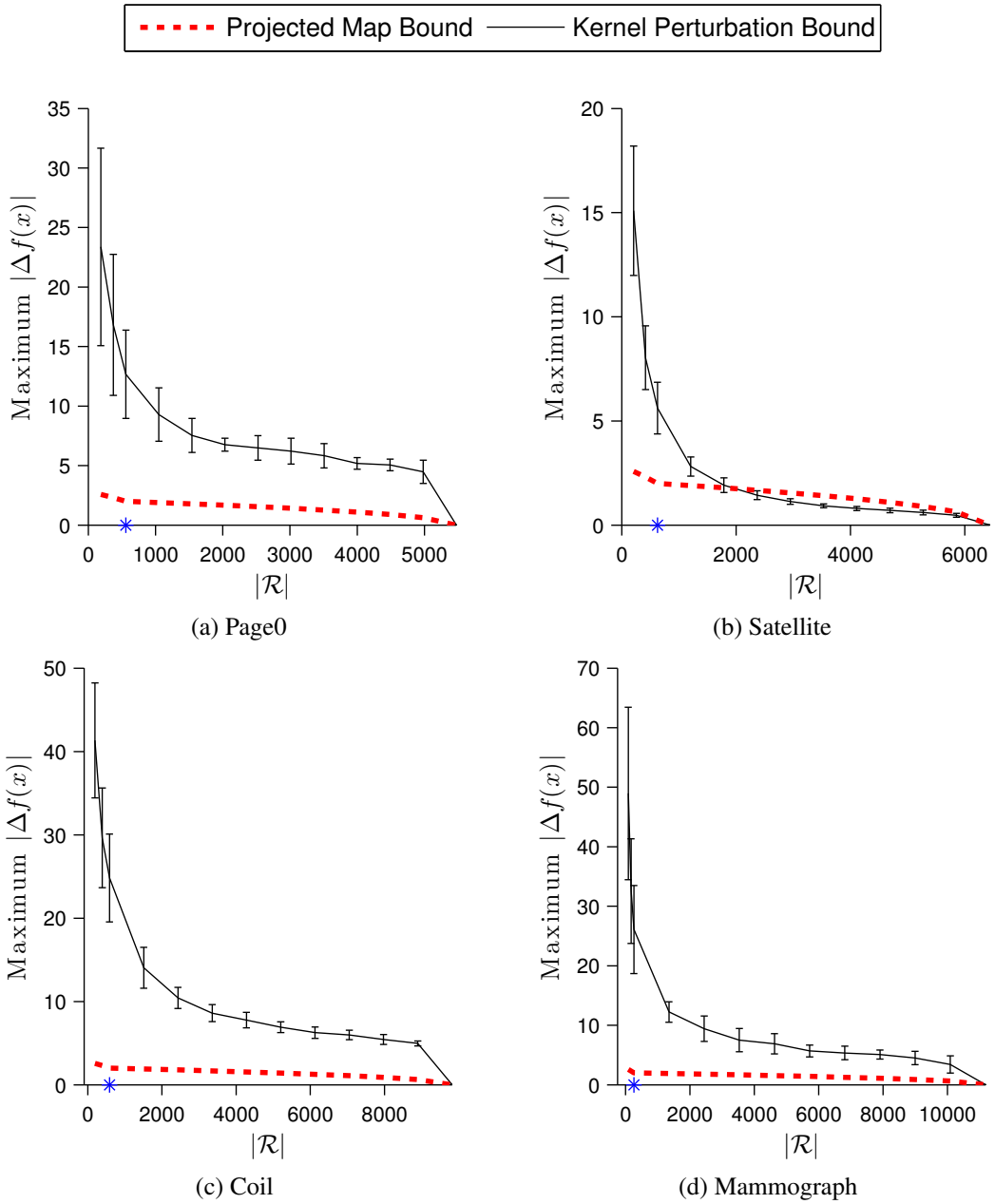


Figure 3: Comparison of the projected map bound (31) (Theorem 3) and the kernel perturbation bound (46) (Theorem 7) as $|\mathcal{R}|$ is increased. The projected map bound is computed assuming rare class points are used in the hypothesis first. The kernel perturbation bound is obtained by randomly sampling a set of basis functions 40 times. The mean value of the bound is plotted and standard deviation is shown as error bars. The blue '*' on the x-axis indicates the number of rare (positive) examples in the dataset.

Figure 3 compares bound (46) with the projected map bound (31) obtained in Theorem 3 as

Name	Source	Subject	d	m	m_+	ρ
Page0	Keel	Computer	10	5472	559	10.2%
Satellite	UCI	Nature	36	6435	626	9.7%
Coil	KDD	Business	85	9822	586	6.0%
Mammograph	(Woods et al., 1993)	Life	6	11183	260	2.3%

Table 1: List of datasets and their characteristics used in Figure 3. d is the number of features, m is the total number of observations, m_+ is the number of rare class observations, and $\rho = \frac{m_+}{m}$ is the percentage of rare class examples.

$|\mathcal{R}|$ is increased on four unbalanced datasets described in Table 1. For the setup, we assume a Gaussian kernel, with width $\sigma^2 = \frac{1}{m^2}\sum_{i,j=1}^{m} ||\mathbf{x}_i - \mathbf{x}_j||_2^2$. Since we are using a Gaussian kernel, $\kappa = 1$. We set $\lambda = 1$, since it does not affect the comparison. For bound (31), we assume kernel functions corresponding to rare class points are included in the representation first. This leads to a deterministic trajectory as $|\mathcal{R}|$ is increased for each dataset. For bound (46), we randomly sample $|\mathcal{R}|$ columns 40 times. For each sample we compute K' to be used in (46). We show the mean and standard deviation of the bound for each value of $|\mathcal{R}|$. From Figure 3 it is clear that the projected map based bound can be significantly lower than the kernel perturbation bound, particularly when $|\mathcal{R}| \ll m$.

Note, the kernel perturbation bound (46) does not depend on class label information. To minimize (46), we need to minimize $||K' - K||_2$, where K' is a Nyström approximation. We can approach this using any one of the various strategies available in the literature for landmark selection in the Nyström method (e.g see Smola and Schölkopf, 2000; Zhang et al., 2008; Farahat et al., 2011). However, better landmark selection is generally achieved at the expense of higher space and time costs. As a result uniform random sampling without replacement remains the method most commonly used in practice (Kumar et al., 2009).

In contrast, the projected map bound (31) uses class label information and captures the asymmetry associated with an unbalanced RankSVM problem. The result leads to a simple selection strategy: to include kernel functions corresponding to the rare class points first. This is compatible with the motivation presented in Section 2.2 for RankRC. Computational results for RankRC, presented in Tayal et al. (2013), confirm that the rare class representation performs better than an equal number of randomly selected points for unbalanced ranking problems.

Finally, one could consider selection methods that combine both insights. For example, for big

datasets, we can select $|\mathcal{R}| < m_+$ points from the positive class using a more sophisticated landmark selection strategy than random sampling. In this case, the extra selection expense may be more acceptable, since we are restricting ourselves to a smaller set of columns, $m_+ \ll m$. On the other hand, if we are interested in selecting $|\mathcal{R}| > m_+$ kernel functions for the hypothesis representation, we can first select the rare class points, and then randomly sample the remaining $|\mathcal{R}| - m_+$ points from the majority class.

5. Predicting Days in Hospital: A Multi-Level Rare Class Ranking Example

In many prediction problems, labels correspond to a set of more than two ordered categories or levels. This situation is referred to as ordinal regression (e.g. see Herbrich et al., 2000). If samples from one of the levels are plenty, while samples from the other levels are rare, then the problem can be considered a multi-level rare class problem. In this section, we extend the biclass RankRC algorithm to handle multiple levels and apply it to a large-scale health informatics problem with a skewed distribution.

The motivating application is based on a recent competition sponsored by the Heritage Health Provider Network (HPN, Accessed: 2013-08-31).[3] The objective is to predict the number of days, $y_i \in \{0, 1, ..., 14, 15+\}$, member i will be hospitalized (inpatient or emergency room visit) in the following year using historical claims data. The number of days a member spends in the hospital is capped at 15 days to help protect the identiy of patients. The data provided consists of three years of historical member claims information. Claims data is anonymized to protect the identity of members (El Emam et al., 2012). The raw data contains basic member information, claims data, drug counts, lab counts and outcome data in a set of relational tables. The training data consists of 147 473 patients over a two year period for which outcomes are given, with on average 12 claims per patient per year (1 764 561 total claims). The third year of data is used for testing, for which outcome information is not provided. We extracted 441 features from the relational data for each patient.

Figure 4 shows the outcome distribution for the two years of training data. We see that the distribution is highly skewed. In particular, examples corresponding to $y_i = 0$ constitute the majority

[3] The competition ran for over two years ending in April 2013 and was highly publicized due to the potential impact on US healthcare and a US $3Mil prize. The authors participated in the competition placing 4th out of 1600+ teams. Our final submission used additional dataset variants and results from other methods as well, which were combined using a model stacking approach.

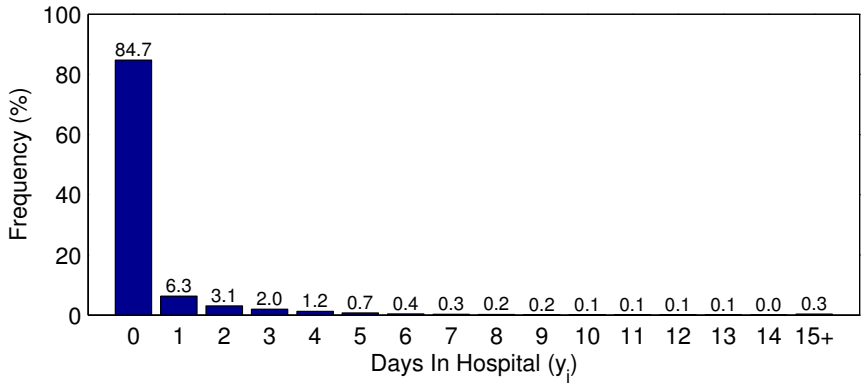


Figure 4: Distribution of the days in hospital for the Heritage Health Network problem.

of cases (85%), while other outcomes, $y_i = 1, \ldots, 15$, are significantly fewer. As in most rare class problems, we are more interested in identifying these rare outcomes.

To solve this problem, one may use traditional metric regression or multi-class classification approaches. However, neither of these approaches correctly capture the structure encoded in the labels. Traditional regression models assume the labels form an interval scale and errors of the same interval are penalized equally. But in the hospitalization prediction problem, errors are not all equal. For example, it is more important to distinguish between 0 and 1 days of hospitalization than between 14 and 15 days. Moreover, it is unclear what transformation would be most appropriate to represent the levels. Consequently, a regression approach may lead to a biased model with poor generalization ability (refer to Herbrich et al., 2000, for further discussion). On the other hand, these levels are also different from the labels of multiple classes in classification problems due to the existence of ordering information.

Therefore, this setting is best handled using an ordinal regression or ranking loss function, which attempts to rank the levels in the correct order, while not depending on the representation of the ranks (Herbrich et al., 2000; Chapelle and Keerthi, 2010). The biclass RankSVM problem (5) introduced in Section 2.1, can be generalized to multiple levels for this purpose, as follows:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^m} \quad \frac{1}{\sum_{r<s} m_r m_s} \sum_{r=1}^{R} \sum_{\{i:y_i=r\}} \sum_{\{j:y_i>y_j\}} \ell_h\left(f(\mathbf{x}_i) - f(\mathbf{x}_j)\right) + \frac{\lambda}{2} \sum_{i,j=1}^{m} \beta_i \beta_j k(\mathbf{x}_i, \mathbf{x}_j)\,, \tag{47}$$

where the hypothesis

$$f(\mathbf{x}) = \sum_{i=1}^{m} \beta_i k(\mathbf{x}_i, \mathbf{x}) \ ,$$

uses kernel instances at all data points following the Representer Theorem. Compared to (5), the loss function in (47), includes an additional summation over each rank level, $r > 0$, with a maximum rank of R. We assume the rank index starts at 0. For each r value, the objective in (47) reduces to a biclass ranking problem with r as the positive class label and all examples with label less than r as the negative class. Thus Problem (47) can be seen as combining R separate biclass ranking problems using the same hypothesis. The constant, m_r, denotes the number of observations which have output value r. For the hospitalization prediction problem, we set $R = 15$, and $r = 0, ..., 15$ represents the different ordinal levels (i.e. days in hospital).

We extend the notion of a rare-class representation to multiple levels as follows. In Figure 4 we observe that $r = 0$ correspond to the majority class, while all other outcomes represent rare cases. Therefore, we consider a representation which only uses kernel functions from examples corresponding to $r = 1, ..., 15$. If we decomposed the problem into R separate binary rare-class problems, this set would constitute the union of all the rare class points used in each of the problems. Thus we obtain the following multi-level RankRC problem:

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^m} \quad \frac{1}{\sum_{r<s} m_r m_s} \sum_{r=1}^{R} \sum_{\{i:y_i=r\}} \sum_{\{j:y_i>y_j\}} \ell_h\left(\bar{f}(\mathbf{x}_i) - \bar{f}(\mathbf{x}_j)\right) + \frac{\lambda}{2} \sum_{\substack{\{i:y_i \neq 0\} \\ \{j:y_j \neq 0\}}} \beta_i \beta_j k(\mathbf{x}_i, \mathbf{x}_j) \ , \tag{48}$$

where the hypothesis

$$\bar{f}(\mathbf{x}) = \sum_{\{i:y_i \neq 0\}} \beta_i k(\mathbf{x}_i, \mathbf{x}) \ ,$$

is constrained to the set of rare class kernel functions. Problem (48) can be solved using similar method as described in Tayal et al. (2013) for the biclass RankRC, by noting the gradient and Hessian of the loss function is simply the sum of R biclass ranking loss functions. Thus, the complexity is $O\left(\sum_{r<s} m_r m_s\right)$ in both space and time.

To evaluate models we count the number of pairs that are correctly ranked among all possible pairs of data objects:

$$\text{MAUC} = \frac{1}{\sum_{r<s} m_r m_s} \sum_{r=1}^{R} \sum_{\{i:y_i=r\}} \sum_{\{j:y_i>y_j\}} \mathbb{I}\left(f(\mathbf{x}_i) > f(\mathbf{x}_j)\right) \ .$$

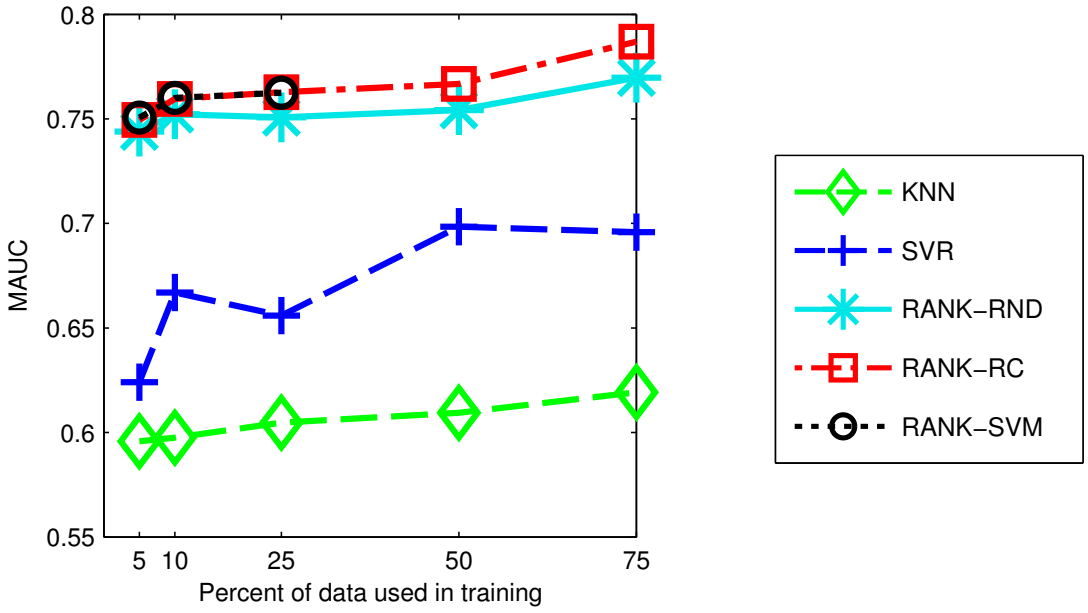
We call this measure MAUC to denote Multi-level AUC. An alternative measure is volume under the ROC surface, which generalizes ROC analysis to ordinal regression. However, computing volume under surface is prohibitive since it has exponential complexity in the number of ordinal levels. MAUC is an approximation of the volume under surface, which can be computed efficiently (Waegeman et al., 2006).

For our experiment, we compare the following methods: k-Neareast Neighbor (KNN) regression, Support Vector Regression (SVR) and three multi-level ranking methods, RANK-SVM (47), RANK-RC (48), and RANK-RND. RANK-RND is similar to RANK-RC, but with the hypothesis restricted to a randomly selected set of kernel functions, with the same cardinality used in RANK-RC.

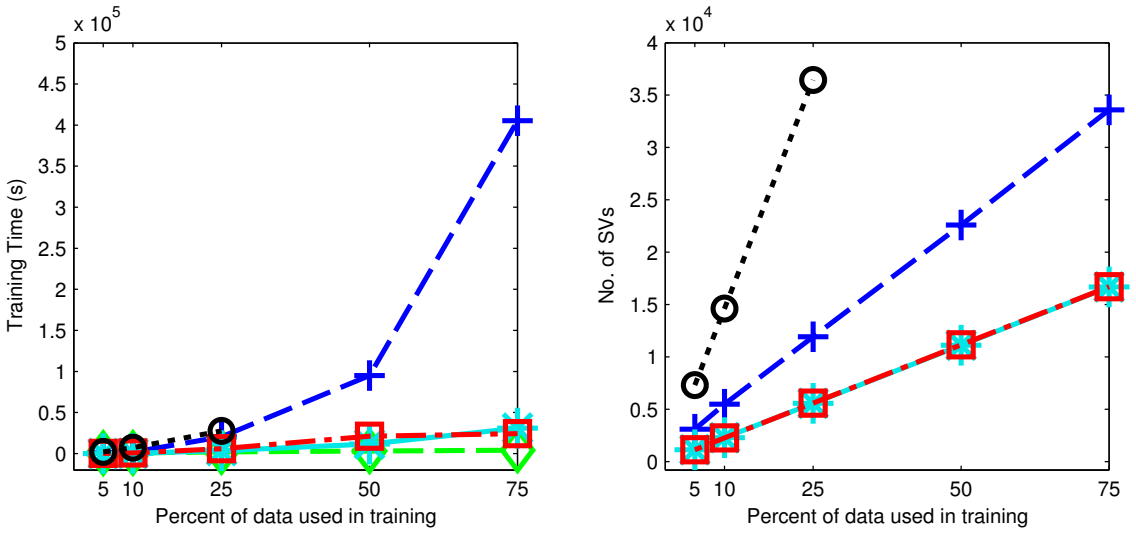
We use LIBSVM (Chang and Lin, 2011) to solve the SVR problem. LIBSVM is a popular and efficient implementation of the sequential minimal optimization algorithm (Platt, 1999). We set cache size to 10GB to minimize cache misses; termination criteria and shrinking heuristics are used in their default settings. The ranking methods (RANK-SVM, RANK-RC, RANK-RND) are solved using the subspace-trust-region method as described in Coleman and Li (1994) and Branch et al. (1999). Termination tolerance is set at 1e-6. For ranking methods, the memory available to store the kernel matrix is limited to 10GB. Experiments are performed on a Xeon E5620@2.4Ghz running Linux.

We train using 5%, 10%, 25%, 50%, and 75% of the training data. Half of the remaining data is used for validation, the other half for test. Features are standardized to zero mean and unit variance before training. Since our focus is on nonlinear kernels, for SVR and the ranking methods, we use the Gaussian kernel, $k(\mathbf{u},\mathbf{v}) = \exp(-\|\mathbf{u}-\mathbf{v}\|_2^2/\sigma^2)$ with $\sigma^2 = \frac{1}{m^2}\sum_{i,j=1}^{m}\|\mathbf{x}_i-\mathbf{x}_j\|_2^2$. The penalty parameter λ (or $\frac{1}{C}$ for SVR) is determined by cross-validation over values $\log_2\lambda = [-20,-18,...,8,10]$. For KNN we cross-validate over $k = [1,2,...,100]$, where k is the number of nearest neighbors.

Figure 5a shows test MAUC results as training data is increased. Note, we are unable to train RANK-SVM with more than 25% of the data as the kernel matrix no longer fits in memory. The ranking methods outperform KNN and SVR. Among the ranking methods, RANK-RC performs slightly better than RANK-RND, and produces almost identical results to RANK-SVM in the cases where RANK-SVM can be computed. Figures 5b and 5c compare training time and number of support vectors, respectively, as training data is increased. We observe RANK-RC and RANK-RND scale well and use fewer support vectors than SVR and RANK-SVM.



(a)



(b) (c)

Figure 5: Comparison of (a) test MAUC (see text) score, (b) training time in seconds, and (c) number of support vectors, for the Heritage Health Network problem as percent of data used for training is increased from 5% to 75%. In our experiment setup, we were unable to train RANK-SVM with more than 25% of the data, due to the large size of the dataset.

6. Conclusion

Many practical data mining problems, such as patient hospitalization, fraud detection, or customer churn share a common characteristic: the cases we are most interested in are in the minority or rare class. Standard algorithms are unable to learn rare class concepts well, since rare class examples are under represented in the dataset sample. In addition, with growing amounts of data, we continually face larger and larger datasets. Recently, Tayal et al. (2013) propose a solution to address challenges associated with large-scale rare class learning called RankRC. Like RankSVM, RankRC is a kernel method that minimizes ranking loss, while learning a regularized optimal hypothesis function. Minimizing ranking loss corresponds to maximizing the AUC for a biclass problem, which is more suitable for rare class datasets than a classification loss. In addition, RankRC exploits class imbalance to achieve computational efficiency by enforcing a rare class hypothesis representation.

In this paper, we analyze the solution of RankRC and compare it to the solution of RankSVM, which uses the complete set of training data for the hypothesis representation. More generally, we consider an arbitrary loss minimization problem, and examine the effect of restricting the hypothesis to any subset of kernel functions ($\mathcal{R}$-subset representation). We show that restricting the hypothesis to a $\mathcal{R}$-subset representation is equivalent to using a projected feature map while solving the unrestricted problem. We use this result to conduct a stability analysis of the $\mathcal{R}$-subset hypothesis for RankSVM. The resulting bound is proportional to $\sqrt{p_+ + p_-}$ where, p_+ is the percentage of points in the positive (rare) class and not in $\mathcal{R}$ and p_- is the percentage of the points in the negative (majority) class but not in $\mathcal{R}$. Therefore, for a fixed cardinality $|\mathcal{R}|$, this bound is minimized by including as many rare class points in the $\mathcal{R}$-subset representation as possible. This result provides further theoretical justification for the RankRC algorithm proposed in Tayal et al. (2013).

In addition, we show that using a $\mathcal{R}$-subset representation is equivalent to solving the original regularized loss minimization problem with a Nyström approximation of the kernel matrix. The Nyström approximation is formed using columns indexed by the set $\mathcal{R}$. This implies that RankRC can be considered as a special Nyström approximation method for RankSVM, with columns selected from the rare class only. Another implication is that we can obtain stability bounds for the $\mathcal{R}$-subset representation using a kernel perturbation approach. However, bounds obtained using the kernel perturbation approach for a Nyström approximation can be arbitrary large. In contrast,

the analysis using the projected feature map approach leads to more stable and tighter bounds. We illustrate this behavior computationally, by comparing bounds obtained for the RankSVM $\mathcal{R}$-subset classifier using the two different approaches.

Although our motivation has been to analyze RankRC, we note that the results we obtain on the equivalency of using a $\mathcal{R}$-subset classifier, a projected feature map, and a Nyström kernel approximation are quite general. These relationships can be used to analyze and devise algorithms for other approximate kernel problems as well.

Finally, in this paper we also extend the biclass RankRC problem to a ranking problem with more than two levels. Our motivating example is based on a competition problem proposed by the Heritage Health Provide Network to predict number of days a member will be hospitalized in the following year. Since the training data contains almost 150 000 samples, the kernel RankSVM problem is too large to solve on standard machines. However, since the outcome distribution is highly skewed, we are able to take advantage of the rare class representation to efficiently solve the problem, with no apparent degradation in performance.

References

Heritage Provider Network Health Prize. `http://www.heritagehealthprize.com/c/hhp`, Accessed: 2013-08-31.

D. Achlioptas, F. McSherry, and B. Schölkopf. Sampling Techniques for Kernel Methods. In *NIPS*, volume 14, pages 335–342, 2001.

C. T. H. Baker. *The numerical treatment of integral equations*. Clarendon Press, 1977.

P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, Classification, and Risk Bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard. A study of the behavior of several methods for balancing machine learning training data. *SIGKDD Explor. Newsl.*, 6(1):20–29, 2004.

B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, COLT '92, pages 144–152. ACM, 1992.

O. Bousquet and A. Elisseeff. Stability and generalization. *J. Mach. Learn. Res.*, 2:499–526, 2002.

A. P. Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30:1145–1159, 1997.

M. A. Branch, T. F. Coleman, and Y. Li. A subspace, interior, and conjugate gradient method for large-scale bound-constrained minimization problems. *SIAM J. Scientific Computing*, 21(1): 1–23, 1999.

C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/ cjlin/libsvm`.

O. Chapelle. Training a support vector machine in the primal. *Neural Comput.*, 19(5):1155–1178, 2007.

O. Chapelle and S. S. Keerthi. Efficient algorithms for ranking with svms. *Inf. Retr.*, 13(3): 201–215, 2010.

N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

N. V. Chawla, N. Japkowicz, and A. Kotcz. Editorial: special issue on learning from imbalanced data sets. *SIGKDD Explor. Newsl.*, 6(1):1–6, 2004.

T. F. Coleman and Y. Li. On the convergence of interior-reflective newton methods for nonlinear minimization subject to bounds. *Mathematical programming*, 67(1-3):189–224, 1994.

C. Cortes, M. Mohri, and A. Talwalkar. On the impact of kernel approximation on learning accuracy. In *Conference on Artificial Intelligence and Statistics*, 2010.

E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics*, 44(3):837–845, 1988.

K. El Emam, L. Arbuckle, G. Koru, B. Eze, L. Gaudette, E. Neri, S. Rose, J. Howard, and J. Gluck. De-identification methods for open health data: the case of the heritage health prize claims dataset. *J Med Internet Res*, 14(1), 2012.

K. Ezawa, M. Singh, and S. W. Norton. Learning goal oriented bayesian networks for telecommunications risk management. In *ICML*, pages 139–147, 1996.

A. K. Farahat, A. Ghodsi, and M. S. Kamel. A novel greedy algorithm for nyström approximation. In *AISTATS*, pages 269–277, 2011.

S. Fine and K. Scheinberg. Efficient svm training using low-rank kernel representations. *J. Mach. Learn. Res.*, 2:243–264, 2002.

C. Fowlkes, S. Belongie, F. Chung, and J. Malik. Spectral grouping using the nyström method. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(2):214–225, 2004.

J. A. Hanley and B. J. Mcneil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, 1982.

H. He and E. A. Garcia. Learning from imbalanced data. *IEEE Trans. on Knowl. and Data Eng.*, 21(9):1263–1284, 2009.

R. Herbrich, T. Graepel, and K. Obermayer. *Large Margin Rank Boundaries for Ordinal Regression*. MIT Press, 2000.

N. Japkowicz and S. Stephen. The class imbalance problem: A systematic study. *Intell. Data*

Anal., 6(5):429–449, 2002.

T. Joachims. Optimizing search engines using clickthrough data. In *KDD*, pages 133–142, 2002.

G. Karakoulas and J. Shawe-Taylor. Optimizing classifiers for imbalanced training sets. In *NIPS*, pages 253–259, 1999.

G. Kimeldorf and G. Wahba. A correspondence between Bayesian estimation of stochastic processes and smoothing by splines. *Ann. Math. Statist.*, 41:495–502, 1970.

M. Kubat and S. Matwin. Addressing the curse of imbalanced training sets: One-sided selection. In *ICML*, pages 179–186, 1997.

S. Kumar, M. Mohri, and A. Talwalkar. Sampling techniques for the nyström method. In *AISTATS*, pages 304–311, 2009.

Y. Lin, Y. Lee, and G. Wahba. Support vector machines for classification in nonstandard situations. *Machine Learning*, pages 191–202, 2000.

M. A. Maloof. Learning when data sets are imbalanced and when costs are unequal and unknown. In *ICML*, 2003.

C. E. Metz. Basic principles of ROC analysis. *Seminars in nuclear medicine*, 8(4):283–298, 1978.

X. Nguyen, M. J. Wainwright, and M. I. Jordan. On surrogate loss functions and f-divergences. *Annals of Statistics*, 37(2):876–904, 2009.

J. C. Platt. Advances in kernel methods. chapter Fast training of support vector machines using sequential minimal optimization, pages 185–208. MIT Press, 1999.

J. C. Platt. Fastmap, metricmap, and landmark mds are all nystrom algorithms. In *In Proceedings of 10th International Workshop on Artificial Intelligence and Statistics*, pages 261–268, 2005.

F. Provost, T. Fawcett, and R. Kohavi. The case against accuracy estimation for comparing induction algorithms. In *In Proceedings of the Fifteenth International Conference on Machine Learning*, pages 445–453, 1997.

B. Raskutti and A. Kowalczyk. Extreme re-balancing for svms: a case study. *SIGKDD Explor. Newsl.*, 6(1):60–69, 2004.

S. Rosset, J. Zhu, and T. Hastie. Margin maximizing loss functions. In *Advances in Neural Information Processing Systems (NIPS 15)*. MIT Press, 2003.

B. Schölkopf, R. Herbrich, and A. J. Smola. A generalized representer theorem. In *Proceedings of the 14th Annual Conference on Computational Learning Theory and and 5th European Conference on Computational Learning Theory*, pages 416–426, 2001.

A. J. Smola and B. Schölkopf. Sparse greedy matrix approximation for machine learning. In *ICML*, pages 911–918, 2000.

Y. Sun, M. S. Kamel, A. K. C. Wong, and Y. Wang. Cost-sensitive boosting for classification of imbalanced data. *Pattern Recogn.*, 40(12):3358–3378, 2007.

A. Talwalkar. *Matrix Approximation for Large-scale Learning*. PhD thesis, Courant Institute of Mathematical Sciences, New York University, New York, NY, 2010.

A. Tayal, T. F. Coleman, and Y. Li. Rankrc: Large-scale nonlinear rare class ranking. *IEEE Transactions on Knowkedge and Data Engineering*, submitted, 2013.

P. D. Turney. Types of cost in inductive concept learning. In *ICML*, 2000.

W. Waegeman, B. D. Baets, and L. Boullart. A comparison of different roc measures for ordinal regression. In *ICML*, 2006.

G. M. Weiss. Mining with rarity: a unifying framework. *SIGKDD Explor. Newsl.*, 6(1):7–19, 2004.

C. Williams and M. Seeger. Using the nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems 13*, pages 682–688. MIT Press, 2001.

K. Woods, C. Doss, K. Bowyer, J. Solka, C. Preibe, and P. Keglmyer. Comparative evaluation of pattern recognition techniques for detection of microcalcifications in mammography. *International Journal of Pattern Recognition and Artificial Intelligence*, 7:1417–1436, 1993.

G. Wu and E. Y. Chang. Class-boundary alignment for imbalanced dataset learning. In *ICML*, pages 49–56, 2003.

J. Zhang and I. Mani. KNN Approach to Unbalanced Data Distributions: A Case Study Involving Information Extraction. In *ICML*, 2003.

K. Zhang, I. W. Tsang, and J. T. Kwok. Improved nyström low-rank approximation and error analysis. In *Proceedings of the 25th international conference on Machine learning*, ICML '08, pages 1232–1239, 2008.

K. Zhang, L. Lan, Z. Wang, and F. Moerchen. Scaling up kernel svm on limited resources: A low-rank linearization approach. pages 1425–1434, 2012.

M. Zhu, W. Su, and H. A. Chipman. LAGO: A Computationally Efficient Approach for Statistical Detection. *Technometrics*, 48, 2006.