Ranking with Distance Metric Learning for Biomedical Severity Detection

Feiyu Xiong, Moshe Kam, Leonid Hrebien
Department of ECE
Drexel University
Philadelphia, PA 19104
fx26@drexel.edu, kam@minerva.ece.drexel.edu, lhrebien@coe.drexel.edu

Yanjun Qi
Department of Computer Science
University of Virginia
Charlottesville, VA 22904
yanjun@virginia.edu

ABSTRACT

Estimating the severity of disease states or adverse-reactions to treatments is very important in drug and therapy development. We have developed a data-driven approach that uses the known severity of both negative controls (least severe) and positive controls (most severe) to define the range of possible severity and used this to learn a distance metric from data. This metric is used to measure the distances of an unknown disease or reaction from both the negative controls and positive controls and thus to estimate its severity. We evaluated three known data sets which studied the severity of fetal hypoxia and toxic reactions of chemical compounds using our approach as well as other four approaches. The results showed that our approach was better able to correctly estimate the severity of the disease/reaction whereas regression based approaches or using other distance metric was much less robust in estimating the corrected results.

Keywords: distance metric learning, severity estimation, biomedical data mining

1. BACKGROUND AND MOTIVATION

Severity estimation is the assessment of the levels of disease states or adverse-reactions to a treatment (drug, regiment, behavior modification, etc.). There are many reasons to study severity estimation. For example, physicians would like to evaluate the stage of a condition/disease so that they can match a treatment to the severity at which a condition is manifested; and track the progression of a condition/disease. Researchers have developed expert based diagnostic scores for tracking disease states and predicting clinical outcomes [3], however, the process is time consuming and expensive [14]. Therefore, researchers try to apply different machine learning algorithms to biomedical problems to develop efficient and effective ways to estimate the severity of diseases or adverse treatment reactions [14][13]. For example, Shankle et al. used decision tree and naive bayes classifiers to identify dementia severity [13]. Tsanas et al. developed a rapid and accurate method for Parkinson's disease severity assessment using speech signal processing and machine learning [14]. These methods were specifically designed for evaluating the severity levels of particular disease. In this paper we focus on developing a general approach for estimating the severity levels of certain biomedical conditions. The phrase "biomedical condition" or "condition" represents a type of disease, or an adverse treatment reaction in the rest of this paper.

The problem of estimating the condition's severity can be divided into two stages: (1) choosing the most medically relevant set of features describing the condition of interest and (2) combining these variables in a functional form (model) which is able to provide the most accurate severity estimation for the condition [3]. Focusing on the stage (2), this paper proposes to tackle Biomedical Severity Estimation using Distance Metric Learning (SE-DML). More specifically, we have observed that in most cases of biomedical severity estimation in practice, the reference data (i.e. the sample groups with known severity) normally includes only positive (e.g. least severe disease state) and negative controls (e.g. most severe disease state). This is because in biomedical experiments such as blood assay, clinical trials and animal testing, many researchers utilize and label positive and negative controls to verify the success of their experiments. Thus SE-DML aims to solve the following problem:

- We are given a data set with multiple samples groups associated with different severity levels of a biomedical condition. Some sample groups' severity levels are known (positive and negative control groups) and some are unknown. Our main goal is to estimate the severity of unknown sample groups based on their relationship to the known ones.

Samples in the same group should match to the same level of severity. For example, a "group" could describe a certain disease state. In order to evaluate the SE-DML approach, we used the following three data sets: one data set for the severity of fetal hypoxia states based on Cardiotocograph (CTG) [1]; two data sets for the severity of toxic activities of chemical compounds - Pyrimidines and Triazines - based on Qualitative Structure Activity Relationship (QSAR) [10]. We also compare four other approaches for estimating the severity of these three data sets. All the results indicate that our approach provided the best overall performance.

2. FORMULATION OF THE TASK

2.1 Problem Definition

Our basic setup includes a data set of m-dimensional samples about a certain biomedical condition. These samples belong to n sample groups $\{\mathbf{E}_1, \dots, \mathbf{E}_n\}$, where each sample group $\mathbf{E} \in \mathbb{R}^{p_E \times m}$ contains p_E samples $\{\mathbf{x}_1, \dots, \mathbf{x}_{p_E}\}$ and corresponds to a severity level y_i of this biomedical condition. We assume that the severity levels y_i are numerical

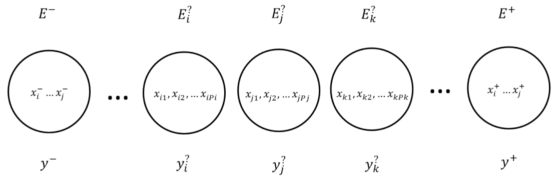

Figure 1: Problem Definition: the severity levels of positive control $\mathbf{E}^+$ and negative control $\mathbf{E}^-$ are known. The severity level $y_i^?$ of an unknown sample group $\mathbf{E}_i^?$ is estimated based on its distances to the two controls.

values between 0 and 1, with 0 being the least severe and 1 being the most severe. Among these n sample groups, some have known severity levels. As we mentioned above, in most cases, the sample groups with known severity are positive and negative controls. Here we define $y^+ = 1$ for positive control $\mathbf{E}^+$, whereas $y^- = 0$ for negative control $\mathbf{E}^-$. The objective is to estimate the severity level $y_i^?$ of an unknown sample group $\mathbf{E}_i^?$ based on its distances to $\mathbf{E}^+$ and $\mathbf{E}^-$. The problem definition of SE-DML is illustrated in Figure 1.

2.2 Connection to Distance Metric Learning

Learning a good distance metric in feature space is critical in machine learning. Several studies show how well simple nearest neighbor methods work if an appropriate distance measure is chosen [9]. Clustering algorithms such as k-means also rely on the pairwise distance measurements between examples [16].

Given a set of k samples $\mathbf{x}_1, , \dots, \mathbf{x}_k$, each $\mathbf{x}_i \in \mathbb{R}^m$ is a data vector with m features. Most metric learning methods try to learn a Mahalanobis distance defined in Equation 1, where A is a positive semi-definite m by m matrix learned from data. The learning process is usually based on the original feature representations and some extra side information which is often available in the form of pairwise constrains on the data: (1) equivalence constraints (Equation 2), which state that the given pair are semantically-similar and should be close together in the learning metric. and (2) inequivalent constraints (Equation 3), which indicate that the given points are semantically-dissimilar and should not be near in the learned metric [17].

$$d_A(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^T A (\mathbf{x}_i - \mathbf{x}_j)} \tag{1}$$

$$\mathbf{S} = \{(\mathbf{x}_i, \mathbf{x}_j) | \mathbf{x}_i \textit{ and } \mathbf{x}_j \textit{ are similar}\} \tag{2}$$

$$\mathbf{D} = \{(\mathbf{x}_i, \mathbf{x}_j) | \mathbf{x}_i \textit{ and } \mathbf{x}_j \textit{ are dissimilar}\} \tag{3}$$

The most commonly used formulation for distance metric learning converts the above constraints to the following convex programming problem [16] for learning the parameter matrix A:

$$\begin{aligned} \min_{A \in \mathbb{R}^{m \times m}} \quad & \sum_{\mathbf{x}_i, \mathbf{x}_j \in \mathbf{S}} d_A(\mathbf{x}_i, \mathbf{x}_j) \qquad (4) \\ \text{s.t.} \quad & \sum_{\mathbf{x}_i, \mathbf{x}_j \in \mathbf{D}} d_A(\mathbf{x}_i, \mathbf{x}_j) \geq 1, \\ & A \succeq 0. \end{aligned}$$

For our targeted task handeling a set of sample groups mapping to a range of severity levels, it is natural to think that one can calculate the distances between samples with unknown severity to samples with known severity, in order to estimate the unknown severity. But the commonly used Euclidean distance metric may not capture the fact that samples from the positive control E^+ should be far from samples from the negative control E^-. The basic idea of distance metric learning is maximizing the distances between dissimilar sample groups, and minimizing the distances between samples in the same group or among similar groups. Specifically, the learned metric based on positive control and negative control should give a maximum distance $d(E^+, E^-)$ between these two controls. The distances between a group $E^?$ of samples with unknown severity level and two controls can then be measured based on this learned metric. These distances should be proportional to $d(E^+, E^-)$ and can be combined to locate the position of this unknown group between the two controls, where the position indicates the severity level.

3. SE-DML APPROACH

3.1 Overall Framework

The objective of SE-DML approach is to estimate the severity levels of n unknown sample group $\{\mathbf{E}_1^?, \dots, \mathbf{E}_n^?\}$ based on positive control $\mathbf{E}^+$ and negative control $\mathbf{E}^-$, which are known before hand. The set of equivalence constrains S (Equation 2) consists of pairs of samples within $\mathbf{E}^+$ or $\mathbf{E}^-$. The set of inequivalent constrains D (Equation 3) consists of pairs of samples from different controls - one sample from $\mathbf{E}^+$ and one sample from $\mathbf{E}^-$. A Mahalanobis distance metric is then learned based on these constrains using the distance metric learning method described in Section 3.2. Based on the learned metric, the distances of the unknown groups to the controls are calculated and will be transformed to severity levels $\mathbf{y}$ as described in Section 3.3.

3.2 Information-Theoretic Metric Learning

We use Information-Theoretic Metric Learning (IMTL) as the metric learning algorithm in our approach since ITML is fast and scalable [7][8]. This algorithm solves the metric learning problem as minimizing the relative entropy between two multivariate Gaussians under side constraints. The formulation of ITML is similar as described in Section 2.2. Two samples are similar if the Mahalanobis distance between them is smaller than a given upper bound, i.e., $d_A(\mathbf{x}_i, \mathbf{x}_j) \leq u$ for a relatively small value of u. Similarly, two samples are dissimilar if $d_A(\mathbf{x}_i, \mathbf{x}_j) \geq l$ for a relatively large l. The objective is to learn a Mahalanobis distance parameterized by A which should be as close as possible to a prior distance function A_0, e.g. Euclidean distance. The closeness of the the solution to the prior is measured by the Kullback-Leibler (KL) divergence [11]:

$$\mathrm{KL}(p(\mathbf{x}; A_0) || p(\mathbf{x}; A)) = \int p(\mathbf{x}; A_0) \log \frac{p(\mathbf{x}; A_0)}{p(\mathbf{x}; A)} d\mathbf{x}. \tag{5}$$

Given pairs of similar points in S and pairs of dissimilar points in D, ITML is defined as is

$$\begin{aligned} \min_{A} \quad & \mathrm{KL}(p(\mathbf{x}; A_0) || p(\mathbf{x}; A)) \qquad (6) \\ \text{s.t.} \quad & d_A(\mathbf{x}_i, \mathbf{x}_j) \leq u, (i, j) \in S, \\ & d_A(\mathbf{x}_i, \mathbf{x}_j) \geq l, (i, j) \in D, \\ & A \succeq 0. \end{aligned}$$

To solve this optimization problem, ITML repeatedly computes Bregman projections [4], which are the projections of the current solution onto a single constraint via the following update

$$A_{t+1} = A_t + \beta A_t(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T A_t, \tag{7}$$

where β is the projection parameter involves the constrains label and step size.

3.3 Severity Estimation for a Sample Group

After the distance metric learned, we can calculate the distances between a sample $\mathbf{x}_i^?$ (with an unknown severity level $y_{x_i}^?$) to the positive control $\mathbf{E}^+$ and negative control $\mathbf{E}^-$. Thus the distance between $\mathbf{x}_i^?$ and $\mathbf{E}^+$ is defined as :

$$d_A(\mathbf{x}_i^?, \mathbf{E}^+) = (\frac{\sum\limits_{x_k \in \mathbf{E}^+} \mathbf{x}_k^+}{|\mathbf{E}^+|} - \mathbf{x}_i^?)^T A (\frac{\sum\limits_{x_k \in \mathbf{E}^+} \mathbf{x}_k^+}{|\mathbf{E}^+|} - \mathbf{x}_i^?). \tag{8}$$

Similarly, the distance between $\mathbf{x}_i^?$ and $\mathbf{E}^-$ is defined as :

$$d_A(\mathbf{x}_i^?, \mathbf{E}^-) = (\frac{\sum\limits_{x_k \in \mathbf{E}^-} \mathbf{x}_k^-}{|\mathbf{E}^-|} - \mathbf{x}_i^?)^T A (\frac{\sum\limits_{x_k \in \mathbf{E}^-} \mathbf{x}_k^-}{|\mathbf{E}^-|} - \mathbf{x}_i^?). \tag{9}$$

These two distances are used together to determine the severity level $y_{x_i}^?$ (Equation 10) for a sample $\mathbf{x}_i^?$. If $y_{x_i}^?$ is closer to 0, the severity level of $\mathbf{x}_i^?$ is more similar to that of the negative control. If $y_{x_i}^?$ is close to 1, the severity level of $\mathbf{x}_i^?$ is more similar to that of the positive control.

$$y_{x_i}^? = \frac{d_A(\mathbf{x}_i^?, \mathbf{E}^-)}{(d_A(\mathbf{x}_i^?, \mathbf{E}^+) + d_A(\mathbf{x}_i^?, \mathbf{E}^-))}. \tag{10}$$

The severity $y_i^?$ of $\mathbf{E}_i^?$ is then defined as

$$y_i^? = \frac{\sum\limits_{\mathbf{x}_i^? \in \mathbf{E}_i^?} y_{x_i}^?}{|\mathbf{E}_i^?|}. \tag{11}$$

4. EXPERIMENT

4.1 Data Sets

The three data sets used in the evaluation are CTG, Pyrimidines and Triazines data sets. Cardiotocography (CTG) is the most widely used tool for fetal surveillance. CTG records the changes in the fetal heart rate and their temporal relationship to uterine contractions [1]. One aim of CTG is to identify fetuses that may be short of oxygen (hypoxic) and then implement corrective treatment plans. In this experiment, we use SE-DML approach to estimate the severity of fetal hypoxic states using a data set from UCI data repository [2][1].

Structure-activity relationships relate to the interaction of chemical compounds with biological systems. These relationships are essential to toxicological investigation in the development of pharmaceutical compounds. Qualitative Structure Activity Relationship (QSAR) is a computer-based modeling method to predict and characterize chemical toxicity [10]. Our severity estimation approach is used to predict the toxicity of two families of chemical compounds, Pyrimidines and Triazines, based on their QSAR data sets [6]

4.2 Experimental Setup

The descriptions of the three data sets are shown in Table 1, including their severity estimation target, severity levels (number of sample groups), sample size and attribute size. The experimental setup is shown in Table 2. Since there are only 3 classes in CTG data set, 90% of the normal samples are used as negative control $\mathbf{E}^-$ and 90% of the pathologic samples are used as positive control $\mathbf{E}^+$. These controls are used to learn distance metric. The remaining 10% of both normal samples and pathologic samples, and all the suspect samples, are used as test classes to evaluate the learned metric. This process is repeated 10 times. The final results are average of the 10 iterations.

For Pyrimidines and Triazines data sets, we randomly picked samples from each severity level. Samples from level 1 and level 5 are used as negative control $\mathbf{E}^-$ and positive control $\mathbf{E}^+$, respectively. The three middle sample groups $\{\mathbf{E}_1^?, \mathbf{E}_2^?, \mathbf{E}_3^?\}$ are used as test groups. The final results are the average of the 20 times random sampling.

For all three data sets, the constrained sample pairs used in our approach are formulated by the samples within negative controls $\mathbf{E}^-$ and positive controls $\mathbf{E}^+$. The lower and upper bounds of the right hand side of the constraint (l and u) in Equation 6 are the 5^{th} and 95^{th} percentiles of the observed distribution of distances between pair of points within each data set.

We implement the following five approaches to compare their ability of estimating the severity on three data sets:

- SE-DML where we use ITML as the metric learning algorithm;
- Euclidean distance under the same framework of SE-DML;
- Large Margin Nearest Neighbors (LMNN), another state-of-the-art metric learning algorithm [15]. We use LMNN under the same framework of SE-DML;
- Linear Regression where we use $\mathbf{E}^+$ and $\mathbf{E}^-$ with severity level 1 and 0, respectively, to build the regression model to predict severity levels of individual samples in each test class $\{\mathbf{E}_1^?$, $\mathbf{E}_2^?$ and $\mathbf{E}_3^?\}$;
- Support Vector Regression, using the same setup as linear regression, implemented by libsvm v3.18 with radial basis function kernel function [5].

We use two evaluation criteria for comparing these five approaches. The first is the relative orders of average severity levels of each test group. Second, in order to measure how well each sample's estimated severity level lies within its group, we use silhouette coefficient [12], which contrasts the average distance to other samples in the same cluster with the average distance to samples in the other clusters. Silhouette coefficient has a value between -1 and 1 where a higher value indicates that the sample is well-matched to its own group, and poorly-matched to the other groups.

Table 1: Descriptions of Three Data Sets

Data Sets	Severity Estimation Target	Severity Levels	Sample Size	Attribute Size
CTG	Severity of fetal hypoxia states	3	2126	21
Pyrimidines	Severity of toxic activities of Pyrimidines	5	74	27
Triazines	Severity of toxic activities of Triazines	5	186	60

Table 2: Experimental Setup for Three Data Sets

	Groups of Different Severity Levels				
Data Sets	$\mathbf{E}^-$	$\mathbf{E}_1^?$	$\mathbf{E}_2^?$	$\mathbf{E}_3^?$	$\mathbf{E}^+$
CTG	90% Normal	10% Normal	100% Suspect	10% Pathologic	90% Pathologic
Pyrimidines	Level 1	Level 2	Level 3	Level 4	Level 5
Triazines	Level 1	Level 2	Level 3	Level 4	Level 5

4.3 Results

The estimated severity levels of the three data sets by the 5 approaches are shown in Figure 2. The three bar charts (row-wise) represent results of the three data sets. In each chart, there are five sets of bars representing the severity levels (y_1, y_2 and y_3) of the three test groups $\mathbf{E}_1^?$, $\mathbf{E}_2^?$ and $\mathbf{E}_3^?$ estimated by the 5 approaches. The blue bar is the severity level y_1 of $\mathbf{E}_1^?$, the green bar is the severity levels y_2 of $\mathbf{E}_2^?$, and the red bar is the severity levels y_3 of $\mathbf{E}_3^?$. Each y_i is the mean of sample severity levels within $\mathbf{E}_i^?$. The standard deviation is shown as the error bar in the figure.

For each set of bars, we ignore the absolute difference between y_1, y_2 and y_3 but only evaluate relative order of these three severity levels. We consider the relative order of $y_1 < y_2 < y_3$ as the correct severity estimation since it matches the true group label. For CTG data set, all the 5 approaches correctly estimate the relative ordering among the severity levels of the three test groups. For pyrimidines data set, linear regression fails to distinguished the severity levels of $\mathbf{E}_1^?$ and $\mathbf{E}_2^?$ and the results show a large standard deviation, indicating it can not robustly estimate the severity of pyrimidines data set. For triazines data set, SE-DML approach is the only one can correctly identify the relative order of $y_1 < y_2 < y_3$.

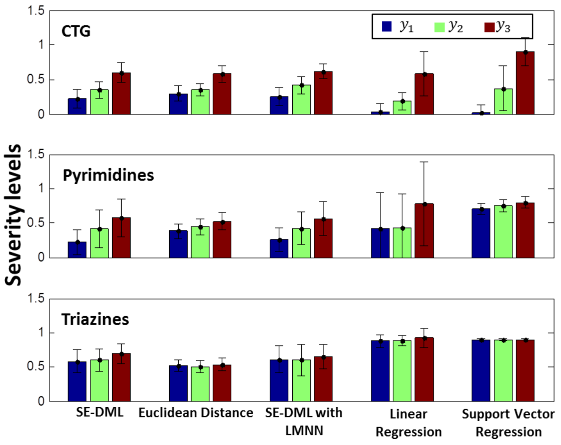

Figure 2: Severity estimation results of the 5 approaches on three data sets. Each bar chart presents the estimated severity levels of one data set.

According to the relative orders among the predicted group-level severity levels, our SE-DML approach has achieved the best performance among all the 5 approaches. However, this could not capture the sample-level severity estimation since only using average severity level representing a group of samples may not be convincing enough. Thus we need to evaluate how well the individual sample's severity level matches to other samples within its group. Silhouette coefficient provides a numerical measure about this evaluation.

For each data set, there are three test groups giving three clusters of individual sample's severity levels. Higher silhouette coefficient of a sample indicates its severity level is well-matched to its own group, when compared to severity levels of samples in other groups. The average silhouette coefficient of the 5 approaches for 3 data sets are listed in Table 3. The bold number in each row indicates the best silhouette coefficient of each data set. SE-DML approach has the best silhouette coefficient in 2 out of the 3 data sets.

Table 3: Silhouette Coefficients of the 5 approaches

	SETC	**SETC with Euclidean Distance**	**SETC with LMNN**	**Linear Regression**	**Support Vector Regression**
CTG	0.2233	-0.0271	0.2987	0.2905	**0.6657**
Pyrimidines	**0.1293**	0.0063	0.0905	-0.0753	0.0116
Triazines	**0.1186**	-0.0135	0.0302	-0.2814	0.1128

5. CONCLUSION

In this paper, we propose an SE-DML approach to estimate the severity levels of biomedical conditions through distance metric learning. This approach first uses samples from positive and negative controls to learn the distance metric which can accurately describe the agreement between unknown samples' distance to controls and their severity levels. The distances of unknown sample to the negative controls and positive controls are calculated based on this learned metric. Then the severity levels of unknown sample are determined by the aggregated ranking score of these distances. Through three experiments, we demonstrated that the proposed approach correctly estimate the severity levels of unknown sample groups.

6. REFERENCES

[1] Ayres-de Campos, D., Bernardes, J., Garrido, A., Marques-de Sá, J., and Pereira-Leite, L. Sisporto 2.0: A program for automated analysis of cardiotocograms. *The Journal of Maternal-Fetal Medicine 9*, 5 (2000), 311–318.

[2] Bache, K., and Lichman, M. UCI machine learning repository, 2013.

[3] Birkner, M. D., Kalantri, S., Solao, V., Badam, P., et al, Pai, M., and Hubbard, A. E. Creating diagnostic scores using data-adaptive regression: An application to prediction of 30-day mortality among stroke victims in a rural hospital in india. *Therapeutics and clinical risk management 3*, 3 (2007), 475–484.

[4] Bregman, L. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *Computational Mathematics and Mathematical Physics 7*, 3 (1967).

[5] Chang, C.-C., and Lin, C.-J. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology 2* (2011), 27:1–27:27.

[6] Chu, W., and Ghahramani, Z. Gaussian processes for ordinal regression. *Journal of Machine Learning Research 6* (2004), 2005.

[7] Davis, J., Kulis, B., Sra, S., and Dhillon, I. Information-theoretic metric learning. In *in NIPS 2006 Workshop on Learning to Compare Examples* (2007).

[8] Davis, J. V., Kulis, B., Jain, P., Sra, S., and Dhillon, I. S. *Information Theoretic Metric Learning.* UT, Austin, http://www.cs.utexas.edu/users/pjain/itml/.

[9] Ding, H., Trajcevski, G., Scheuermann, P., Wang, X., and Keogh, E. Querying and mining of time series data: Experimental comparison of representations and distance measures. *Proc. VLDB Endow. 1*, 2 (Aug. 2008), 1542–1552.

[10] Hansch, C., Hoekman, D., Leo, A., Zhang, L., and Li, P. The expanding role of quantitative structure-activity relationships (qsar) in toxicology. *Toxicology Letters 79*, 1–3 (1995), 45 – 53. Decision Subtances Methodologies for Human Health Risk Assessment of Toxic Substances.

[11] Kullback, S., and Leibler, R. A. On information and sufficiency. *The Annals of Mathematical Statistics 22*, 1 (1951), pp. 79–86.

[12] Rousseeuw, P. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math. 20*, 1 (1987), 53–65.

[13] Shankle, W. R., Mania, S., Dick, M. B., and Pazzani, M. J. Simple models for estimating dementia severity using machine learning. *Studies in health technology and informatics 52 Pt 1* (1998), 472–476.

[14] Tsanas, A., Little, M., McSharry, P., and Ramig, L. Accurate telemonitoring of parkinson's disease progression by noninvasive speech tests. *Biomedical Engineering, IEEE Transactions on 57*, 4 (2010), 884–893.

[15] Weinberger, K. Q., and Saul, L. K. Distance metric learning for large margin nearest neighbor classification. *J. Mach. Learn. Res. 10* (2009), 207–244.

[16] Xing, E., Ng, A., Jordan, M., and Russell, S. *Distance Metric Learning with Application to Clustering with Side-Information.* MIT Press, 2003, pp. 505–512.

[17] Yang, L. Distance metric learning: A comprehensive survey, 2006.