

Learning to Rank the Severity of Unrepaired Cleft Lip Nasal Deformity on 3D Mesh Data

Jia Wu

Department of Electrical Engineering
University of Washington
Seattle, Washington, 98195, U.S.A.
Email: jiauw@uw.edu

Raymond Tse

Seattle Children's Hospital
Department of Surgery
University of Washington
Seattle, WA, 98195, U.S.A.
Email: raymond.tse@seattlechildrens.org

Linda G. Shapiro

Department of Electrical Engineering,
Computer Science and Engineering
University of Washington
Seattle, Washington, 98195, U.S.A.
Email: shapiro@cs.washington.edu

Abstract—Cleft lip is a birth defect that results in deformity of the upper lip and nose. Its severity is widely variable and the results of treatment are influenced by the initial deformity. Objective assessment of severity would help to guide prognosis and treatment. However, most assessments are subjective. The purpose of this study is to develop and test quantitative computer-based methods of measuring cleft lip severity. In this paper, a grid-patch based measurement of symmetry is introduced, with which a computer program learns to rank the severity of cleft lip on 3D meshes of human infant faces. Three computer-based methods to define the midfacial reference plane were compared to two manual methods. Four different symmetry features were calculated based upon these reference planes, and evaluated. The result shows that the rankings predicted by the proposed features were highly correlated with the ranking orders provided by experts that were used as the ground truth.

Keywords—learning to rank; 3D shape quantification; cleft lip; face symmetry.

I. INTRODUCTION

Cleft lip occurs in approximately 1 in 1000 newborn children and can be associated with cleft palate [1]. The deformity is thought to result from a failure of fusion in utero and may be associated with underdevelopment of tissues [2]. Surgical treatment can produce a dramatic change in appearance of the lip; however, stable correction of the nose remains a challenge, and treatment strategies continue to be debated. Given that the potential results of treatment are limited by the cleft severity, objective assessment of the deformity is important for prognosis and treatment outcomes.

Traditional evaluation of cleft deformities was relied on clinical description and landmark-based measurements that are taken directly with calipers. Neither of these is ideal, given that clinical descriptors are somewhat subjective, and anthropometric measurements on young infants are difficult and burdensome.

Advances in 3D stereophotogrammetry have made rapid capture of 3D facial form a practical reality for infants. The accuracy of the indirect anthropometric measurements made on these images has been evaluated [3], however this analysis still relies heavily on manual input. An automated computer-based system for facial analysis would greatly facilitate medical researchers.

Computer-based tools have been developed and used to study autism [4], plagiocephaly [5] and 22q11.2 deletion syndrome [6]. However, none have been used to study infants with unrepaired clefts. Our goal is to develop novel tools for analysis of shape in children with cleft lip. Specifically, we want these tools to be automated, computer-based, and quantitative. Given that facial asymmetry increases directly with increasing cleft severity, this study focuses on quantifying nasolabial symmetry.

In this paper, we present a system that learns to rank the severity of the abnormalities of the 3D infant faces with cleft. Our system uses a midfacial plane as a reference to compute the difference between the left and the right side of the face according to four different features. After the differences are extracted, a machine learning algorithm takes the ranking orders provided by an expert as the ground truth to train a classifier to order the data according to the severity of the clefts.

The rest of the paper is organized as follows: section II describes the dataset used to develop and test the system. In section III, five different methods for computing midfacial planes are introduced, including three automatic approaches and two manual approaches for comparison. In section IV, our grid-patch based symmetry features are described in detail. Section V defines the algorithms that are used for ranking, and Section VI shows the experimental results of our work.

II. DATASETS

Our cleft dataset consists of 3D craniofacial surface meshes obtained from the 3dMD Craniofacial imaging system [7]. The 3D face models obtained were pre-processed and pose-normalized using an automated system [6]. The dataset contains 35 meshes from infants with unrepaired unilateral cleft lip and 5 normal infant controls. In terms of acquiring the ground truth, although rating specific facial features can produce variable results [8], ranking a group of subjects in a side-by-side comparison can be performed reliably [9]. In order to facilitate comparison and ranking of digital 3D images, we developed an interface that allows the user to freely shuffle 3D images and examine them in a side-by-side manner. Subjects were ranked in order of severity of the cleft lip nasal deformity by an expert cleft surgeon to serve as the ground truth.

TABLE I: Brief summary of the 5 approaches for computing midfacial reference plane

Method	Approach	Details of the approach
mirror	Computer	mirror the data and register the mirrored data and original data [10]
a-lmk	Computer	calculated using landmarks from deformable registration [11]
learning	Computer	midfacial reference plane models learned from training data [12]
m-lmk	Manual	calculated using landmarks from medical expert
plane	Manual	directly put on 3D mesh data

III. DEFINING THE MIDFACIAL REFERENCE PLANE

In order to measure symmetry we needed to define a plane across which asymmetry would be measured. In a perfectly symmetric face this plane would be the plane of symmetry and would equally divide the two halves, but in a face with cleft abnormalities, the asymmetry of the nasolabial region will alter the plane of symmetry relative to the other part of the face and it is hard to define a midfacial reference plane. Multiple automatic computer-based approaches have been developed. One method was introduced by Benz *et al.* [10], in which the original data is mirrored at an arbitrary plane. Then the original mesh and the mirrored mesh are registered using the iterated-closest-point algorithm. In our paper, this method is referred as *the mirror method*. In the second method, referred to as *the a-lmk method*, 24 landmarks are automatically located by a deformable registration algorithm from a template mesh to a target mesh, which is initialized by a geometric point detector [11]. After these landmarks are found, the midfacial reference plane is calculated using only the landmarks on the eyes and chin area. The last method is a learning method which takes two steps of processing: landmark-related region detection and midfacial reference plane computation using these regions [12]. It is referred as *the learning method* in this paper. The the learning method and the a-lmk method come from our own previous work in plane finding [12] and in automatic landmark finding [11].

In additional to the three automatic methods, two sets of midfacial reference planes were provided manually by two craniofacial specialists to be used as ground truth for the automatic methods and further comparison. One (*the m-lmk method*) is based on landmarks, in which the medical experts provide the landmarks on cleft patient data, and the midfacial reference plane is calculated using the eyes and chin landmark positions. In the second one (*the plane method*), the midfacial reference plane was drawn directly on the 3D mesh data by an expert. A brief summary of these five methods is given in Table I.

The accuracy of this midfacial reference plane is critical for all of the next steps, as we will show that the performance of different plane detection methods varies using the same features and the same ranking algorithms.

IV. QUANTIFYING THE ASYMMETRY OF THE FACE

Symmetry measures are defined to quantify the difference between the left and right sides of the face as requested by the the medical experts. Using the midfacial reference plane, several different symmetry measures are defined based on grid patches. A grid is placed over the face and it measures the difference between the left side and right side in terms of

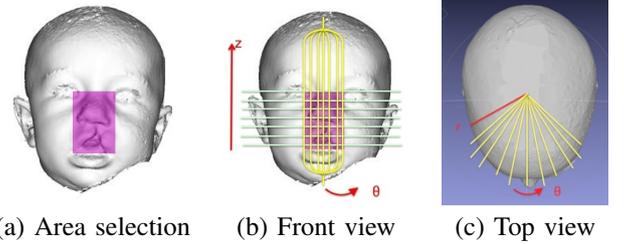


Fig. 1: Selected area, grid patches and r , θ and z directions

features of the corresponding grid patches. The features used are radius difference (from a central point), angle difference, curvature difference, and edge feature difference.

A. Area Selection

Because the deformity mainly occurs in the nose and mouth area, the grid will be placed only in the center part of the face. Using mouth corners and a nose bridge point generated automatically by [11], a rectangle area is cropped from the face as shown in Fig. 1(a).

B. Grid-Patch-Based Quantification

Grid-patch-based quantifications divide the area selected from the face into several patches (as shown in Fig. 1). Each patch is represented by the average value of the points inside it. Half of the rectangle area is divided into M by M squares, equally divided in the z and θ directions, as shown in Fig. 1(b). In our experiments, $M = 10$.

Four differences are compared for each corresponding reflected patch pair: the radius difference (RD), the angle difference (AD), the curvature difference (CD), and the sharp edge difference (ED). The radius difference defined for a grid patch at position (θ, z) is

$$RD(\theta, z) = |r(\theta, z) - r(-\theta, z)|$$

where r takes the average radius value in that grid patch, and $(-\theta, z)$ is the reflected grid patch of (θ, z) with respect to the midfacial reference plane. This gives the actual surface distance. The angle difference defined for grid patch (θ, z) is

$$AD(\theta, z) = \cos(\beta_{v_{\theta,z}, v_{-\theta,z}})$$

where $\beta_{v_{\theta,z}, v_{-\theta,z}}$ is the angle between the surface normal vector of the face mesh at grid patch (θ, z) and its reflected grid patch. This shows how differently the two patches are oriented. For curvature difference, the average Gaussian curvature of

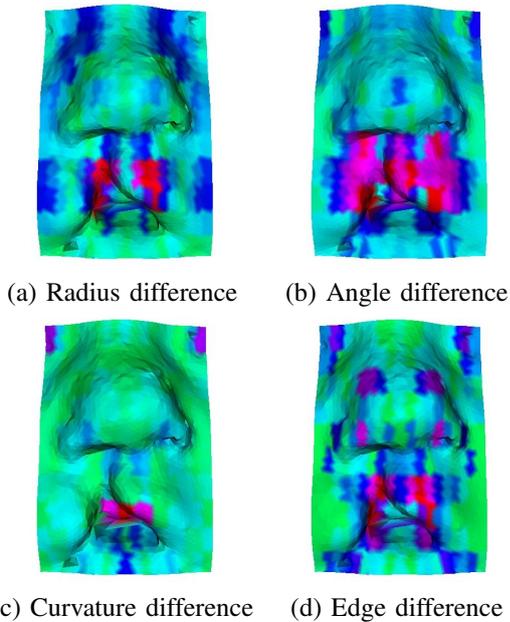


Fig. 2: Grid-patch-based asymmetry measurements. Red means a big difference between that grid patch and its reflecting patch. Green means a small difference.

each grid patch is calculated, represented as $K(\theta, z)$. The Gaussian curvature difference is

$$CD(\theta, z) = |K(\theta, z) - K(-\theta, z)|$$

for grid patch (θ, z) , where $K(-\theta, z)$ is the average Gaussian curvature value for the reflected grid patch. Last but not least, the sharp edge for one grid patch is defined as the ratio of points with a dihedral angle larger than a certain threshold angle Th to the total number of points in the grid patch. The shape angle difference is defined as:

$$ED(\theta, z) = \left| \frac{\#points(\theta, z) > Th}{\#points(\theta, z)} - \frac{\#points(-\theta, z) > Th}{\#points(-\theta, z)} \right|$$

representing, in two corresponding reflected grid patches, the difference between the ratio of points with sharp edges.

Given the local differences of RD , AD , CD , and ED , a vector of length $M \times M$ is formed to represent the asymmetry of the center part of the face. Figure 2 illustrates the four local differences with a 10×10 grid.

V. LEARNING TO RANK

The task of learning from the expert’s ranking to compare the severity of face abnormalities falls into a learning-to-rank problem, which has been studied heavily with applications in many information retrieval problems, such as document retrieval, collaborative filtering, and computational advertising. Liu [13] categorized the algorithms to train a rank model into three groups: pointwise, pairwise, and listwise. In our paper, we will compare two pointwise and two pairwise algorithms.

In our paper, linear regression [14] and SVM regression [15] are used as the pointwise algorithms, and Rank-

Boost [16] and RankNet [17] are used as the pairwise algorithms to learn how to rank the data. The pointwise methods approximate the problem as a regression problem: given a single instance, predict its score. The pairwise algorithms take the features of every pair instance and transfer the problem into a classification problem: learning a binary classifier that can tell which instance is better (higher rank) in a given pair of instances.

RankBoost trains the model in rounds. It starts with all pairs being assigned an equal weight. At each round, the learner selects the weak ranker that achieves the smallest pairwise loss on the training data with respect to the current weight distribution. Pairs that are correctly ranked have their weights decreased and those that are incorrectly ranked have their weights increased so that the learner will focus more on the hard samples in the next round. The final model is essentially a linear combination of weak rankers. Weak rankers theoretically can be of any type but they are most commonly chosen as binary functions with a single feature and a threshold [18].

RankNet is a probabilistic pairwise ranking framework based on neural networks. For every pair that is correctly ranked, each instance is propagated through the net separately. The difference between the two outputs is mapped to a probability by the logistic function. The cross entropy loss is then computed from that probability and the true label for that pair. Next, all weights in the network are updated using the error back propagation and the gradient descent method [18].

VI. EXPERIMENTS AND RESULTS

A. Ground truth

The dataset with unrepaired unilateral cleft lip includes left and right clefts. For better comparison, all the individuals with left cleft lip were mirrored by the plane given by the expert, so they appear to be right cleft in the pictures and 3D meshes. Thus, 35 right cleft meshes and 5 control meshes were shown to the expert. A user interface was created to allow the user to arrange 3D meshes. The expert can click and drag the pictures in any order desired, and also open up a window with 3 neighboring meshes to rotate, enlarge and compare the details in the 3 meshes to carefully determine the order. After the expert is finished ranking, the images have an assigned ground truth rank from 1 to 40.

B. Measurements

Because the ground truth and the scores predicted are all ranks instead of actual quantified linear scores, Spearman rank correlation coefficient ρ and the Kendall rank correlation coefficient τ are used to evaluate the experimental results.

The Spearman correlation coefficient is defined as the Pearson correlation coefficient between the ranked variables. For a sample of size n , the n raw scores X_i, Y_i are converted to ranks x_i, y_i , and ρ is computed from these. The closer the ρ value is to 1, the better the two ranks are correlated:

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

TABLE II: Ranking correlations for all features (feature length 400, CV4). Each box contains Spearman correlation coefficient ρ followed by Kendall correlation coefficient τ .

Method	Linear R	SVM R	RankNet	RankBoost
mirror	0.661 0.522	0.636 0.511	0.512 0.389	0.683 0.515
a-lmk	0.597 0.489	0.599 0.489	0.513 0.389	0.773 0.615
learning	0.574 0.482	0.589 0.515	0.669 0.541	0.746 0.582
m-lmk	0.560 0.478	0.549 0.452	0.632 0.485	0.635 0.493
plane	0.524 0.422	0.521 0.400	0.630 0.533	0.771 0.615

The Kendall τ test is a non-parametric hypothesis test for statistical dependence based on the τ coefficient. It is a pairwise error that reflects how many pairs are ranked discordant. The best matching ranks get a τ value of 1.

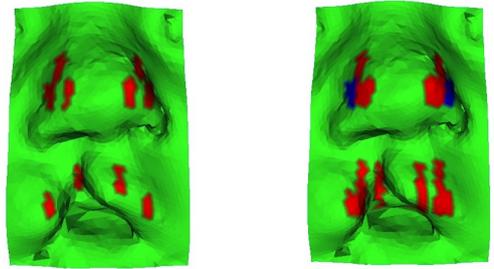
$$\tau = \frac{(\# \text{ concordant pairs}) - (\# \text{ discordant pairs})}{\frac{1}{2}n(n-1)}$$

C. Experiments

For our experiments, we used the WEKA [19] implementation for linear and SVM regression. RankLib [20] was used for RankNet and RankBoost training and testing. For RankNet and RankBoost, every mesh in the test data set is paired with another test mesh, and the rank for each test mesh can be calculated based on this pairwise comparison. The feature length is 400, with 10×10 grids and 4 scores in each grid. 4-fold cross validation was performed, and the results are in Table II, with the first number as the Spearman correlation coefficient ρ and the second number as the Kendall correlation coefficient τ . A small dataset (40 instances) with large feature vectors (400 features) does not perform very well with the regression methods and RankNet. The pairwise method RankBoost performs significantly better than all the other regression and ranking methods. The reason is that RankBoost used weak rankers in every round to pick up a feature that is most distinguishable. In this experiment, it used 10 features in these weak rankers, instead of all 400. In terms of the midfacial reference plane finding methods, the plane produced by the a-lmk method has the highest performance of the three automatic methods, with a score similar to that obtained by the manual plane method in which the experts actually drew the plane on the data.

D. Feature Selection and Results

Based on the good performance accomplished by RankBoost and the reasons behind it, a feature selection was done by best-first search to come up with the most distinguishing features in all 400 features. Out of the top five features selected by the best-first search approach, three of them are angle differences, one is an edge difference, and one is a curvature difference. The grid positions of the top five grid patches are shown in Fig. 3(a). The grid patches are located on the side of the nose area and upper mouth area, which are exactly the areas the experts are looking at when ranking. The top 10 features are shown in Fig. 3(b). The blue-colored grid patch near the nose side is selected twice, with one edge difference and one angle difference. The others contain four angle features,



(a) Top 5 selected grids (b) Top 10 selected grids

Fig. 3: Top discriminative features. (a) The red colored areas are the positions for top 5 selected grid patches. (b) The red and blue areas are the top 10 selected grid patches. The red grid patches are selected once, and the blue patch is selected twice with two features.

TABLE III: Ranking correlations for selected features (feature length 5 CV4)

Method	Linear R	SVM R	RankNet	RankBoost
mirror	0.729 0.589	0.730 0.574	0.719 0.570	0.687 0.528
a-lmk	0.787 0.641	0.780 0.633	0.809 0.663	0.707 0.559
learning	0.792 0.637	0.812 0.644	0.843 0.700	0.750 0.612
m-lmk	0.800 0.648	0.813 0.652	0.831 0.696	0.772 0.611
plane	0.795 0.659	0.813 0.670	0.827 0.711	0.752 0.626

two edge features, one radius difference, and one curvature difference.

Using only the top five features for 4-fold cross validation on the dataset and repeating the same experiments as in VI-C, the results are boosted by 0.06 for ρ (from around 0.77 to more than 0.83), and 0.1 for τ (from around 0.61 to 0.71), as shown in Table III. Out of all four ranking algorithms, RankNet now obtains the best performance. The reason is that RankNet is based on neural networks which are known to be hard to train. When it is dealing with a large number of features, it is less effective than RankBoost [18]. However, when using less features, because the task is much easier, RankNet is trained more efficiently. Out of the three automatic midfacial reference plane finding methods, the learning method is now the winner and is able to beat both manual methods by a small margin. Both the learning method and the a-lmk method beat the mirror method by a substantial margin. A subset of the results is shown in Fig. 4. Ten images with nose and cleft areas are shown from nine unilateral cleft infants and one control. They are ordered by the expert's rank. The other ranks for each midfacial plane method are predicted by RankNet.

VII. CONCLUSION

This paper introduced a system to learn and rank the severity of abnormalities on 3D faces with cleft lip. The system takes a midfacial reference plane, extracts symmetry measurements based on grid patches determined by that midfacial reference plane, and uses a machine learning algorithm to train a model to predict the ranks of how difficult it is to repair the cleft lip nasal deformity. For the midfacial reference plane,

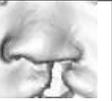
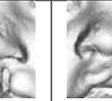
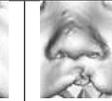
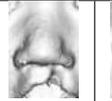
expert's order	1	2	3	4	5	6	7	8	9	10
images										
learning	1	3	2	4	5	6	8	9	7	10
a-lmk	1	2	3	5	6	4	8	7	9	10
mirror	1	2	4	8	5	6	9	3	7	10
m-lmk	1	2	3	4	5	6	9	7	10	8
plane	1	2	3	5	4	6	7	9	10	8

Fig. 4: Ranking results: ten sample images with nose and cleft areas are shown, nine with unilateral cleft and one control. They are ordered by the expert's rank. Under the images are their ranks by our system (ranked 1-10, with 1 being most severe).

three automated methods along with two manual approaches were compared. Four machine learning algorithms were tested to learn from the experts' ranking orders and to build a model to predict the severity of clefts based on the midfacial plane and four patch-based features. The learning method for plane construction along with the algorithm RankNet performs the best. The results show that the rankings predicted by the proposed features are highly correlated with the clinicians' ranking order.

Further studies will include extracting other features related to the severity of clefts beside symmetry features, train and test all the features on a bigger and more complete dataset and apply the method to evaluate the surgery outcomes.

ACKNOWLEDGMENT

This research was supported by NIH/NIDCR under grant number 1U01DE020050-01 (PI: L. Shapiro)

REFERENCES

- [1] H.H. Ardinger, K.H. Buetow, G.I. Bell, J. Bardach, D.R. VanDemark, and J.C. Murray. Association of genetic variation of the transforming growth factor-alpha gene with cleft lip and palate. *American journal of human genetics*, 45(3):348, 1989.
- [2] H. Kalter and J. Warkany. Experimental production of congenital malformations in strains of inbred mice by maternal treatment with hypervitaminosis a. *The American Journal of Pathology*, 38(1):1, 1961.
- [3] R. Tse, L. Booth, K Keys, B. Saltzman, E. Stuhau, Kapadia, H, and C.L. Heike. Reliability of nasolabial anthropometric measures using 3d photogrammetry in infants with unrepaired unilateral cleft lip. *Accepted for publication in Plastic and reconstructive surgery*, 2013.
- [4] P. Hammond, C. Forster-Gibson, AE Chudley, JE Allanson, TJ Hutton, SA Farrell, J. McKenzie, JJA Holden, and MES Lewis. Face-brain asymmetry in autism spectrum disorders. *Molecular psychiatry*, 13(6):614–623, 2008.
- [5] I. Atmosukarto, L.G. Shapiro, J.R. Starr, C.L. Heike, B. Collett, M.L. Cunningham, and M.L. Speltz. Three-dimensional head shape quantification for infants with and without deformational plagiocephaly. *The Cleft Palate-Craniofacial Journal*, 47(4):368–377, 2010.
- [6] K. Wilamowska, L. Shapiro, and C.L. Heike. Classification of 3D face shape in 22q11. 2 deletion syndrome. In *Biomedical Imaging: From Nano to Macro, 2009. ISBI'09. IEEE International Symposium on*, pages 534–537. IEEE, 2009.
- [7] 3dMD. <http://www.3dmd.com>.
- [8] C. Asher-McDade, C. Roberts, W.C. Shaw, and C. Gallager. Development of a method for rating nasolabial appearance in patients with clefts of the lip and palate. *The Cleft Palate-Craniofacial Journal*, 28(4):385–391, 1991.
- [9] D. M. Fisher, R. Tse, and J.R. Marcus. Objective measurements for grading the primary unilateral cleft lip nasal deformity. *Plastic and reconstructive surgery*, 122(3):874–880, 2008.
- [10] M. Benz, X. Laboureux, T. Maier, E. Nkenke, S. Seeger, F.W. Neukam, and G. Häusler. The symmetry of faces. In *Proceedings of Vision, Modeling, and Visualization, G. Girod, H. Niemann, T. Ertl, B. Girod, and H.-P. Seidel, eds.(Akademische Verlagsgesellschaft, 2002)*, pages 43–50, 2002.
- [11] S. Liang, J. Wu, S. M. Weingberg, and L. G. Shapiro. Improved detection of landmarks on 3d human face data. In *IEEE Engineering in Medicine and Biology Society Annual Conference*. IEEE, 2013.
- [12] J. Wu, R. Tse, C.L. Heike, and L.G. Shapiro. Learning to compute the symmetry plane for human faces. In *ACM Conference on Bioinformatics, Computational Biology and Biomedical Informatics*. ACM, 2011.
- [13] Tie-Yan Liu. Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, 3(3):225–331, 2009.
- [14] J. Neter, W. Wasserman, M.H. Kutner, et al. *Applied linear statistical models*, volume 4. Irwin Chicago, 1996.
- [15] S. K. Shevade, S. S. Keerthi, C. Bhattacharyya, and K.R. Murthy. Improvements to the smo algorithm for svm regression. *Neural Networks, IEEE Transactions on*, 11(5):1188–1193, 2000.
- [16] Y. Freund, R. Iyer, R. E. Schapire, and Y. Singer. An efficient boosting algorithm for combining preferences. *The Journal of machine learning research*, 4:933–969, 2003.
- [17] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*, pages 89–96. ACM, 2005.
- [18] V. Dang and B.W. Croft. Feature selection for document ranking using best first search and coordinate ascent. In *Proc. of SIGIR 2010 Workshop on Feature Generation and Selection for Information Retrieval*, 2010.
- [19] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and L.H. Witten. The weka data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, 2009.
- [20] V. Dang. <http://sourceforge.net/p/lemur/wiki/ranklib/>, 2013.