Multi-score Learning for Affect Recognition: the Case of Body Postures

Hongying Meng[1], Andrea Kleinsmith[2], and Nadia Bianchi-Berthouze[1]

[1] UCL Interaction Centre, University College London, London, UK
[2] Goldsmiths, University of London, London, UK
h.meng@ucl.ac.uk, a.kleinsmith@gold.ac.uk, n.berthouze@ucl.ac.uk

Abstract. An important challenge in building automatic affective state recognition systems is establishing the ground truth. When the ground-truth is not available, observers are often used to label training and testing sets. Unfortunately, inter-rater reliability between observers tends to vary from fair to moderate when dealing with naturalistic expressions. Nevertheless, the most common approach used is to label each expression with the most frequent label assigned by the observers to that expression. In this paper, we propose a general pattern recognition framework that takes into account the variability between observers for automatic affect recognition. This leads to what we term a multi-score learning problem in which a single expression is associated with multiple values representing the scores of each available emotion label. We also propose several performance measurements and pattern recognition methods for this framework, and report the experimental results obtained when testing and comparing these methods on two affective posture datasets.

Keywords: Automatic emotion recognition, observer variability, affective computing, affective posture, pattern recognition, multi-labeling, multi-score learning

1 Introduction

With the emergence of the affective computing field [17], various studies have been carried out to create systems that can recognize the affective states of their users by analyzing their vocal, facial [15] [23], and body expressions [9] and even their physiological changes [11]. An important challenge in building such automatic affective state recognition systems is establishing the ground truth, i.e., to label the training and testing sets necessary to build such systems. When the ground truth is not available, researchers recur to the use of perceptual studies where observers are asked to name the affective state conveyed by an expression (e.g., a body expression) and then use the most frequent label to label that expression. This approach assumes that a ground truth exists and that the automatic recognition system should behave as the majority of the observers.

As the field distances itself from acted datasets and begins to focus more on naturalistic expressions, unfortunately the subtlety of naturalistic expressions tends to lower the inter-rater reliability to fair or moderate levels [3]. This is particularly true for affective body expressions. For example, Kleinsmith *et al.* [9] used a random repeated sub-sampling method to assign ground truth labels to naturalistic postures according to groups of naïve observers. The results showed an average level of agreement of 67% between observers. This low level of agreement has also been observed for acted body expressions. Camurri *et al.* [2] examined the level to which groups of observers recognized the emotion portrayed in dance motions performed by professional dancers. The dancers' labels were considered the ground truth. 32 non-expert observers were asked to evaluate both the dance expression and its intensity and an average of 56% correct recognition was achieved. Another example is provided by Paterson et al.'s study [16] in which they also examined how well an actor's affective state may be recognized by a group of observers. Actors were motion captured while performing drinking and knocking motions according to 10 affective states. Human observers viewed the motions and judged the emotion displayed in a forced choice experimental design. The results showed that the overall recognition rate across the 10 emotions was a mere 30%.

Given the variability observed in these perception tasks (for both acted and non-acted datasets), it becomes important to take this variability into consideration. This problem has been addressed in more general terms by the machine learning community. Chittaranjan *et al.* [4] proposed to incorporate the annotator's knowledge into a machine learning framework for detecting psychological traits using multimodal data. They used the knowledge provided by the annotators, the annotations themselves and their confidences in the form of weights to estimate final class labels and then used these to train classifiers. Using this approach, the resulting classifiers outperformed classifiers that were trained using the most frequent label approach.

A different approach is taken by Fürnkranz and Hüllermeier [7]. They proposed to learn preferences rather than a set ground truth. In this case each sample is associated with a set of labels and their preferred value. A supervised learning of a ranking function is used to learn the complete ranking of all labels instead of only predicting the most likely label. The results show that even if the aim is only to predict the most preferred label, the learning process gains from taking into account the full ranking information. As the authors acknowledge, the problem with their approach is that it requires a large training dataset. Nicolaou *et al.* [14] treated the same problem as a regression task of predicting multi-dimensional output vectors given a specific set of input features. They proposed a novel output-associative relevance vector machine (RVM) regression framework that augments the traditional RVM regression by being able to learn non-linear input and output dependencies.

In line with these last two works, we propose a general framework for a pattern recognition problem that considers category labeling differences between observers which is described in the following section.

2 Multi-score Learning and Its Measurements

The scenario considered here is a pattern recognition problem in which there are multiple scores on the categories for a single sample. We call this problem multi-score learning. It should be noted that such a problem is different from typical machine learning problems with multiple outputs, such as multi-class learning, multi-label learning and multi-output learning. It is also different from typical regression models. Here, we describe multi-score learning with a detailed formulation of the problem, measurements and possible learning methods.

Let $X \subset R^d$ denote the d dimensional feature space of samples. Every sample $\mathbf{x} \in X$ has a multiple score vector $\mathbf{y}$ over C categories. All of these score vectors create a C dimensional score space $\{\mathbf{y} \in Y\}$. Because the scores have a maximum value, without losing generalization, we can assume that all the scores are within the interval $[0, 1]$, e.g. $Y \subset [0, 1]^C$. For a training dataset $\left\{(\mathbf{x}^i, \mathbf{y}^i), i = 1, 2, \cdots, N\right\}$, in which $\mathbf{x}^i = (x_1^i, x_2^i, \cdots, x_d^i)$ is a d dimensional feature vector and $\mathbf{y}^i = (y_1^i, y_2^i, \cdots, y_C^i)$ is its score, the machine learning task here is to find a function $h(\mathbf{x}) : X \to Y$ to predict the scores

$$\hat{\mathbf{y}} = h(\mathbf{x}) = (\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_C),\ \hat{y}_j \in [0, 1] \tag{1}$$

for a given sample $\mathbf{x} = (x_1, x_2, \cdots, x_d)$.

In order to measure the performance of possible methods for multi-score learning, we define five measurements to compute the similarity between the true and predicted scores for all testing samples $\left\{(\mathbf{y}^i, \hat{\mathbf{y}}^i), i = 1, 2, \cdots, M\right\}$. These measurements give a full comparison between the true and predicted scores by considering their distance, similarity, ranking order and multi-class, multi-label classification performances.

Root Mean Square Error. Root Mean Square Error (RMSE) is a frequently-used measure of the differences between the values predicted by a model and the values actually observed from the object being modeled or estimated. The average RMSE over all the testing samples is computed as

$$\text{RMSE} = \frac{1}{M}\sum_{i=1}^{M}\sqrt{\frac{1}{C}\sum_{j=1}^{C}(y_j^i - \hat{y}_j^i)^2} \tag{2}$$

Cosine Similarity. Cosine Similarity is a measure of similarity between two vectors by measuring the cosine of the angle between them. The result of the Cosine function is equal to 1 when the angle is 0, and less than 1 when the angle is of any other value.

$$\cos(\theta) = \frac{\langle \mathbf{y}, \hat{\mathbf{y}} \rangle}{\|\mathbf{y}\|_2 \|\hat{\mathbf{y}}\|_2} = \frac{\sqrt{\sum_{j=1}^{C} y_j \hat{y}_j}}{\sqrt{\sum_{j=1}^{C} y_j^2}\sqrt{\sum_{j=1}^{C} \hat{y}_j^2}} \tag{3}$$

The average cosine similarity (ACS) on the testing dataset is computed as

$$\text{ACS} = \frac{1}{M}\sum_{i=1}^{M}\frac{\left\langle \mathbf{y}^i, \hat{\mathbf{y}}^i \right\rangle}{\|\mathbf{y}^i\|_2 \left\|\hat{\mathbf{y}}^i\right\|_2} \tag{4}$$

Top Match Rate. Top match rate (TMR) evaluates how many times the top-ranked label is not the same as the top label of the sample. It is the same as the recognition error for multi-class classification.

$$\mathrm{TMR} = \frac{1}{M}\sum_{i=1}^{M} \mathbf{1}_{\left\{\underset{1\le j\le C}{\mathrm{argmax}}\mathbf{y}_j^i = \underset{1\le j\le C}{\mathrm{argmax}}\hat{\mathbf{y}}_j^i\right\}} \tag{5}$$

where $\mathbf{1}_A$ is a function on condition A.

$$\mathbf{1}_A = \begin{cases} 1, \, A \text{ is true} \\ 0, \, A \text{ is false} \end{cases} \tag{6}$$

Ranking Loss. The order of the predicted scores among C categories might be more important as it gives a relative comparison between these categories. The ranking loss (RL) measure considered here is based on an information retrieval application[19]. RL evaluates the average fraction of label pairs that are reverse ordered for the sample [24]. Assume that for sample $\mathbf{x}^i$, its real score $\mathbf{y}^i$ can be represented in order as $(y^i_{l_1} \ge y^i_{l_2} \ge \cdots \ge y^i_{l_C})$ and a predicted score $\hat{\mathbf{y}}^i$. With this inderstanding, the average RL function can be defined as

$$\mathrm{ARL} = \frac{1}{M}\sum_{i=1}^{M} \frac{\sum_{j=1}^{C}\sum_{k=j+1}^{C} \mathbf{1}_{\left\{\hat{y}^i_{l_j} < \hat{y}^i_{l_k}\right\}}}{C\times(C-1)/2} \tag{7}$$

Average Precision. In order to compare the overall recognition rate for multiple categories, average precision (AP) can be also considered. It is an important measurement for the average recognition rate for a multi-label classification problem. The problem is transferred into a multi-label classification task by thresholding($\ge \delta$) the true label into value "1" and "0". i.e. new labels $\mathbf{y}^i \in \{0,1\}^C$. AP measures the average fraction of labels ranked above a particular label l which has an actual value of "1" (e.g. $\mathbf{y}^i_l = 1$). The performance is perfect when the value is 1.

$$\mathrm{AP} = \frac{1}{M}\sum_{i=1}^{M} \frac{1}{\sum_{l=1}^{C}\mathbf{1}_{\left\{y^i_l=1\right\}}} \sum_{\substack{l=1 \\ y^i_l=1}}^{C} \frac{\sum_{k=1}^{C}\mathbf{1}_{\left\{\hat{y}^i_k \ge \hat{y}^i_l, y^i_k=1\right\}}}{\sum_{k=1}^{C}\mathbf{1}_{\left\{\hat{y}^i_k \ge \hat{y}^i_l\right\}}} \tag{8}$$

3 Learning Methods for Multi-score Learning

There are many classification or regression methods that could be adapted to perform multi-score learning. The most popular methods are considered and applied here.

3.1 K-Nearest Neighbour

K-Nearest Neighbour (KNN) is a lazy learning method for classifying objects based on the closest training examples in the feature space. For sample $\mathbf{x}$, its predicted label $\hat{\mathbf{y}} = (\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_C)$ can be computed as the average of the labels in its K neighbors $N(\mathbf{x}) \subset \{1, 2, \cdots, N\}$ in the N training samples. i.e.

$$\hat{y_j} = \tfrac{1}{K}\textstyle\sum_{k=1}^{K} y_j^{i_k},\ j = 1, 2, \cdots, C, i_k \in N(\mathbf{x}) \tag{9}$$

3.2 Regression

If we assume the dependent variables are independent from each other, we can use the general linear model (GLM), support vector regression (SVR) or partial least squares (PLS) methods.

General Linear Model. GLM [12] is a statistical linear model that has multivariate measurements $\mathbf{y}$. The feature vector $\mathbf{x}$ is usually assumed to follow a multivariate normal distribution. The components of $\mathbf{y}$ are assumed to be independent from each other. GLM is solved independently by solving a normal regression problem for each component .

Support Vector Regression. The SVR algorithm [6] is very similar to the support vector machine (SVM) algorithm, however it treats the data as a regression problem. The model produced by SVR depends only on a subset of the training data, because the cost function for building the model ignores any training data close to the model prediction.

Partial Least Squares. PLS regression is a statistical method that bears some relation to principal components regression. Instead of finding hyperplanes of maximum variance between the response and independent variables, it finds a linear regression model by projecting the predicted variables and the observable variables to a new space. Because both the X and Y spaces are projected to new spaces, PLS methods are known as bilinear factor models. Detailed information on the implementation of these methods can be found in [5] and [18].

3.3 Artificial Neural Networks

Artificial neural networks can be applied for multi-score learning without the assumption of independence between dependence variables. Two of these networks are introduced below.

Radial Basis Neural Network. A radial basis neural network (RBNN) [13] typically has three layers: an input layer, a hidden layer with a non-linear radial basis activation function and a linear output layer. The neurons in the hidden layer contain Gaussian transfer functions whose outputs are inversely proportional to the distance from the center of the neuron.

General Regression Neural Networks. A general regression neural network (GRNN) is a probabilistic neural network proposed by Donald F. Specht [21] in 1991. It needs only a fraction of the training samples that a back-propagation neural network needs [21]. The use of a probabilistic neural network is especially advantageous due to its ability to converge to the underlying function of the data with only a few training samples available. The additional knowledge needed to fit the data in a satisfying way is relatively small and can be achieved without additional input by the user.

3.4 Multi-task Learning

Multi-task learning (MTL) [1] is a method for learning sparse representations shared across multiple tasks. It is based on a novel non-convex regularizer which controls the number of learned features common across the tasks. The algorithm has a simple interpretation: it alternately performs a supervised and an unsupervised step. In the supervised step it learns task-specific functions, and in the unsupervised step it learns common-across-tasks sparse representations for these functions. MTL can be applied to multi-score learning by considering every component in $\mathbf{y}$ as a single task.

4 Multi-score Learning on Affective Posture: Results

Two posture datasets[3] were used to test our approach. Both datasets were collected using motion capture systems. Examples of postures from the two datasets can be seen in Figure 1. The first set contains 108 acted postures and each posture is described by a 24-dimensional feature vector. This vector describes the configuration of the posture in terms of distances between body joints and angles between body segments. Details on the data collection are provided in [10]. The second dataset contains 103 non-acted postures collected in a whole-body computer game scenario. For each posture, a 41-dimensional feature vector describing 3D rotational information for each body joint was extracted. Details on this dataset are provided in [9].

Each posture in both databases was labeled using non-expert observers and forced-choice surveys. For the acted dataset, 70 observers from 3 different cultures were asked to rate each posture in terms of 4 emotion labels (anger, fear, happiness and sadness). For the non-acted database, 8 observers made a series of 5 evaluations on the entire set of postures according to 4 affective state labels (concentrated, frustrated, triumphant and defeated). The results of the posture evaluation surveys for both the acted and non-acted datasets are shown in Figure 1. Each posture is represented by a pie chart showing the frequency of use of each label which was computed as the average over the number of observers who assigned that label to that posture. For details on the labeling process see [10] and [9] respectively. The agreement level for the acted dataset reached an average observer agreement of 85% (Cohen's kappa ranged between 0.75-0.84, i.e., substantial to almost perfect) [8]. The results for the non-acted dataset reached an average observer agreement of 67% (Cohen's kappa ranged between 0.30 to 0.62, i.e., fair to moderate), significantly lower than the acted dataset. For details on the relevance of the posture features see [10][9][20].

For each posture, its feature vector and pie chart were used for training the recognition system. In the testing, only the features vectors were input to the system and a pie-like label was produced for each posture representing the probability of each affective state. All the methods mentioned in Section 3 were

[3] Available at: http://www.ucl.ac.uk/uclic/people/n_berthouze/research.

used to test both the acted and non-acted posture datasets. A 10-fold cross-validation method and normalization were used to keep $\hat{\mathbf{y}} \in [0,1]^C$. $K = 5$ for KNN and $\delta = 0.25$ for AP were chosen for both datasets. The average values for 5 different performance measurements obtained by each method are shown in Table 1. In this Table, a "↓" indicates that smaller values correspond to higher performances, whereas a "↑" indicates that higher values correspond to higher performances. The best performances are shown in bold.

Fig. 1. (i) Postures examples. (ii) and (iii) represent the survey results for the two posture datasets. Each pie chart corresponds to the frequency of use for each affective state label for each posture image according to the observers. The pie charts are grouped according to the most frequent label indicated below each group.

Among these five measurements, TMR(Top Match Rate) can be used for comparison with the top-label approach (i.e., most frequent label approach). Therefore, we also computed the performances of the KNN and SVM methods based on the top-label approach. For the non-acted dataset, the top-label approach for KNN and SVM reached 67.9% and 58.8% corect recognition rates, respectively. 59% was reported in [9] using a back propagation method. In Table 1, using the multi-score learning approach, TMR reached 70.2% with KNN and 69.5% with SVR, showing a clear improvement over the top-label approach and the human-observer agreement level (67%) if considered as a baseline. For the acted database, the top-label approach reached 63.9% and 54.1% with the KNN and SVM methods respectively. In comparison, TMR reached 67% with GRNN for multi-score learning. The results for the top-label approach using a multi-layer perception obtained higher recognition rates with performances between 63% and 77% for each separate culture group of observers [8].

Table 1. Performances for the 7 learning methods and the 5 evaluation measurements on both the acted (top) and non-acted (bottom) posture datasets

Acted Postures	KNN	GLM	SVR	PLS	RBN	GRNN	MTL
Root Mean Square Error (↓)	0.165	0.197	0.189	0.193	0.215	**0.161**	0.205
Cosine Similarity (↑)	**0.862**	0.814	0.836	0.822	0.771	0.852	0.807
Top Match Rate (↑)	0.635	0.601	0.595	0.602	0.522	**0.674**	0.619
Ranking Loss (↓)	0.164	0.239	0.223	0.244	0.287	**0.141**	0.269
Average Precision (↑)	**0.761**	0.687	0.717	0.699	0.663	0.756	0.683
Non-acted Postures	**KNN**	**GLM**	**SVR**	**PLS**	**RBN**	**GRNN**	**MTL**
Root Mean Square Error (↓)	0.141	0.161	**0.140**	**0.140**	0.145	0.143	0.139
Cosine Similarity (↑)	0.885	0.850	**0.892**	0.884	0.876	0.880	0.888
Top Match Rate (↑)	**0.702**	0.612	0.695	0.672	0.669	0.687	0.682
Ranking Loss (↓)	0.176	0.210	0.177	0.195	0.216	**0.152**	0.165
Average Precision (↑)	0.679	0.695	0.690	0.693	**0.734**	0.726	0.669

The other 4 measurements shown in Table 1 provide information on the performances of each method over the distributions of all the affective states

for all postures. For example, AP reached 76% correct recognition which is very close to the top-label approach performance (77%). AP of 76% means that the model can correctly predict 76% of top labels (i.e., the labels that have over 25% agreement between observers) for each posture regardless of whether the posture has only one, or more than one label with high observer agreement. In general, these measurements aim to provide a more complete description of the performance of each method. This framework allows for a more comprehensive evaluation of the methods and their properties with respect to the needs of the modeling problem.

5 Conclusions

Multi-score learning is a very common problem in affective computing applications. Agreement between observers is often not very high especially when dealing with naturalistic subtle expressions. This paper provides a framework for multi-score learning problems that take into account the variability between observers. The output scores provide more comprehensive information than single labels.

Overall, the performances of the various methods were very good and comparable to, if not higher than, the human observers' agreement levels and the top-label approach performances for both the acted and non-acted datasets. Furthermore, even when using TMR only, the results show better performance than the top-label approach for the non-acted posture dataset where the agreement between observers is quite low. Multi-score learning uses the complete label information instead of the majority agreed label to make the prediction more accurate and reliable. For TMR, multi-score learning uses a regression method to perform the classification task in which more detailed label information was used. From a learning method perspective, it can be seen that GRNN reliably obtained good performance on both datasets. The reason is that they not only provide non-linear regressions for each score, but also deal with the possible correlations between different categories.

The approach proposed here is also very general and modality independent. Therefore, it would be interesting to test the same approach for other modalities as well as with a fusion of modalities in cases where they may appear to disagree in the type of emotions they convey (e.g., a facial expression incongruent with its body expression). In this case, instead of the observers agreement, the problem of ground truth becomes one of the agreement between modalities. Furthermore, it would be interesting to investigate the benefits of each method with respect to the type of modality and the level of agreement that the data represent [22].

6 ACKNOWLEDGMENTS

This work was supported by EPSRC grant EP/G043507/1: Pain rehabilitation: E/Motion-based automated coaching.

References

1. A. Argyriou, T. Evgeniou, and M. Pontil. Convex multi-task feature learning. *Machine Learning*, 73:243–272, 2008.
2. A. Camurri, I. Lagerlof, and G. Volpe. Recognizing emotion from dance movement: Comparison of spectator recognition and automated techniques. *International Journal of Human-Computer Studies*, 59(1-2):213–225, 2003.
3. G. Castellano and K. Karpouzis and C. Peters and J.-C. Martin(Eds). Special Issue on Real-Time Affect Analysis and Interpretation: Closing the Affective Loop in Virtual Agents and Robots. *Journal on Multimodal User Interfaces*, 3(1):1–3, 2010.
4. G. Chittaranjan, O. Aran, and D. Gatica-Perez. Exploiting observers judgements for nonverbal group interaction. In *Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition (FG'11)*, Santa Barbara, CA, USA, March 2011.
5. S. Jong. Simpls: An alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 18(3):251–263, 1993.
6. H. Drucker, C. Burges, L. Kaufman, A. Smola, and V. Vapnik. Support vector regression machines. In *Advances in Neural Information Processing Systems 9*, volume 9, pages 155–161, 1997.
7. J. Fürnkranz and E. Hüllermeier. Pairwise preference learning and ranking. In *Proceedings of the 14th European Conference on Machine Learning*, pages 145–156. Springer-Verlag, 2003.
8. A. Kleinsmith. Grounding Affect Recognition on a Low-Level Description of Body Posture. *Ph.D. Thesis.* University College London. 2010.
9. A. Kleinsmith, N. Bianchi-Berthouze, and A. Steed. Automatic recognition of non-acted affective postures. *IEEE Transactions on Systems, Man and Cybernetics, Part B.*, 99: 1–12, 2011.
10. A. Kleinsmith, P.R. de Silva, and N. Bianchi-Berthouze. Cross-cultural differences in recognizing affect from body posture. *Interacting with Computers*, 18(6):1371–1389, 2006.
11. R.L. Mandryk, K.M. Inkpen, and T.W. Calvert. Using psychophysiological techniques to measure user experience with entertainment technologies. *Behaviour & IT*, 25(2):141–158, 2006.
12. K.V. Mardia, J.T. Kent, and J. M. Bibby. *Multivariate Analysis.* Academic Press, 1980.
13. J. Moody and C.J. Darken. Fast learning in networks of locally-tuned processing units. *Neural Comput.*, 1(2):281–294, 1989.
14. M.A. Nicolaou, H. Gunes, and M. Pantic. Output-associative rvm regression for dimensional and continuous emotion prediction. In *Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition (FG'11)*, Santa Barbara, CA, USA, March 2011.
15. M. Pantic and L.J.M. Rothkrantz. Automatic analysis of facial expressions: The state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:1424–1445, 2000.
16. H.M. Paterson, F.E. Pollick, and A.J. Sanford. The role of velocity in affect discrimination. In *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*, pages 756–761. Lawrence Erlbaum Associates, 2001.
17. R.W. Picard. *Affective Computing.* The MIT Press, 1997.

18. R. Rosipal and N. Krämer. Overview and recent advances in partial least squares. In Craig Saunders, Marko Grobelnik, Steve Gunn, and John Shawe-Taylor, editors, *Subspace, Latent Structure and Feature Selection*, volume 3940 of *Lecture Notes in Computer Science*, pages 34–51. Springer Berlin / Heidelberg.
19. G. Salton. Developments in automatic text retrieval. *Science*, 253(5023):pp. 974–980, 1991.
20. P.R. De Silva and N. Bianchi-Berthouze. Modeling human affective postures:an information theoretic characterization of posture features. *Computer Animation and Virtual Worlds*, 15(3-4):269–276, 2004.
21. D. F. Specht. A general regression neural network. *IEEE Transactions on Neural Networks*, 2(6):568–576, November 1991.
22. J. Wagner and E. Andre and F. Lingenfelser and J. Kim and T. Vogt . Exploring Fusion Methods for Multimodal Emotion Recognition with Missing Data. *IEEE Transactions on Affective Computing*, 99:1949–3045, 2011.
23. Z. Zeng, M. Pantic, G.I. Roisman, and T.S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(1):39–58, 2009.
24. M. Zhang and Z. Zhou. ML-KNN: A lazy learning approach to multi-label learning. *Pattern Recognition*, 40(7):2038–2048, July 2007.