Using entropy as a measure of acceptance for multi-label classification

Laurence A. F. Park and Simeon Simoff

School of Computing, Engineering and Mathematics
University of Western Sydney, Australia
{lapark,s.simoff}@uws.edu.au
http://www.scem.uws.edu.au/~lapark

Abstract. Multi-label classifiers allow us to predict the state of a set of responses using a single model. A multi-label model is able to make use of the correlation between the labels to potentially increase the accuracy of its prediction. Critical applications of multi-label classifiers (such as medical diagnoses) require that the system's confidence in prediction also be provided with the multi-label prediction. The specialist then uses the measure of confidence to assess whether to accept the system's prediction. Probabilistic multi-label classification provides a categorical distribution over the set of responses, allowing us to observe the distribution, select the most probable response, and obtain an indication of confidence by the shape of the distribution. In this article, we examine if normalised entropy, a parameter of the probabilistic multi-label response distribution, correlates with the accuracy of the prediction and therefore can be used to gauge confidence in the system's prediction. We found that for all three methods examined on each data set, the accuracy increases for the majority of the observations where the normalised entropy threshold decreases, showing that we can use normalised entropy to gauge a systems confidence, and hence use it as a measure of acceptance.

1 Introduction

Multi-label learning is the process of learning the association of L binary labels $\boldsymbol{y}$ of the response space, to a given point in a explanatory space $\boldsymbol{x}$. A multi-label classifier may have a high dimensional explanatory space $\mathbb{R}^M$, and a high dimensional response space $\mathbb{B}^L$ (where $\mathbb{B}$ is $\{0, 1\}$), depending on the data. Therefore there may be many suitable responses for a given $\boldsymbol{x}$, but only the most likely is provided as the predicted response. A review of multi-label learning is found in [5].

Critical applications of multi-label learning, such as medical diagnoses, military support or political decisions, require that multi-label predictions are accurate. Therefore it is essential that all predictions are paired with a measure of the multi-label system's confidence in its prediction. If the confidence is high, the specialist can accept the systems prediction. If the confidence is low, the specialist will not accept the prediction, but may also examine the cause of the low confidence.

Probabilistic multi-label learning is the process of assigning a categorical distribution over the $\mathbb{B}^L$ space for a given $\boldsymbol{x}$. Each of the 2^L elements $\boldsymbol{y}$ in the space $\mathbb{B}^L$ are assigned a probability. By examining the distribution, we can determine the most likely response to the input $\boldsymbol{x}$, and also examine if other response label sets have high probability, giving us confidence in the response and an indication of the relationship between the labels.

When many labels exist, it is difficult to examine and compare the distribution over the 2^L label combinations. It is also not always simple to determine a system's confidence by observing the response distribution. Therefore it would be useful to summarise the distribution with a parameter that can be used to measure a system's confidence of its prediction.

In this article, we examine if the accuracy of a probabilistic multi-label system's response is correlated to a parameter of the response distribution. We hypothesise that the normalised entropy of the response distribution will provide us with a measure of confidence. The contributions of this article are:

- a discussion on the use of normalised entropy of the response distribution to assess prediction accuracy (Section 3),
- an analysis of the relationship between accuracy and normalised entropy for probabilistic multi-label classification (Section 4), and
- a probabilistic version of the label powerset multi-label classifier (Section 2.3)

The article will proceed as follows: In section 2 we examine the concept of probabilistic multi-label learning and examine three models for computing the joint distribution over the powerset of labels. Section 3 discusses the use of the response distribution to assess the accuracy of the predicted response label set. Finally, section 4 empirically examines the relationship between normalised entropy and accuracy of a response.

2 Probabilistic Multi-label Learning

For a given input space $\mathbb{R}^M$ and a set of L labels l_i, a probabilistic multi-label classifier learns the probability distribution over the powerset of labels $\mathbb{B}^L$. A probabilistic multi-label classifier maps an input vector $\boldsymbol{x} \in \mathbb{R}^M$ to a categorical probability distribution $\theta \in \mathbb{S}^{2^L}$, where $\mathbb{S}^{2^L}$ is the 2^L dimensional simplex. Using the categorical distribution, we can identify the probability of each label set being the correct response to $\boldsymbol{x}$

$$\theta_i = P(\boldsymbol{y}_i|\boldsymbol{x}) \tag{1}$$

where $\boldsymbol{x}$ is the input vector to be classified, $\boldsymbol{y}_i$ is the ith element in the powerset of labels $\mathbb{B}^L$, and θ_i is the probability that $\boldsymbol{y}_i$ is the correct label set of $\boldsymbol{x}$. For example, given the three labels, l_1, l_2 and l_3 and an input vector $\boldsymbol{x}$, a probabilistic multi-label classifier will provide a distribution over the eight elements of the powerset shown in Table 1, where $\sum_{i=1}^{2^L} \theta_i = 1$.

Table 1. The powerset elements y_i of the labels l_1, l_2 and l_3, and the associated probability θ_i of each label set computed by the probabilistic multi-label learner.

y_i	$\{\}$	$\{l_1\}$	$\{l_2\}$	$\{l_3\}$	$\{l_1, l_2\}$	$\{l_1, l_3\}$	$\{l_2, l_3\}$	$\{l_1, l_2, l_3\}$
$P(y_i\|x_j)$	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8

Constructing a probabilistic multi-label classifier is equivalent to modelling the joint distribution of labels, conditioned on the input $\boldsymbol{x}$:

$$P(l_1, l_2, \ldots, l_L|\boldsymbol{x}) = P(\boldsymbol{y}_i|\boldsymbol{x}) \tag{2}$$

We will now examine probabilistic forms of three common multi-label classifiers.

2.1 Probabilistic Binary Relevance

The simplest form of probabilistic multi-label classifier, called Probabilistic Binary Relevance (PBR), treats each of the labels as independent of each other, giving us:

$$P(l_1, l_2, \ldots, l_L|\boldsymbol{x}) = \prod_{i=1}^{L} P(l_i|\boldsymbol{x}) \tag{3}$$

The task is then simplified to learning the probabilities $P(l_i|\boldsymbol{x})$ for each label i. This independence assumption ignores any correlation between labels and so is equivalent to constructing L independent probabilistic binary classifiers.

2.2 Ensemble of Probabilistic Classifier Chains

Rather than assuming independence, the joint probability can be expressed in terms of a product of conditional probabilities:

$$P(l_1, l_2, \ldots, l_L|\boldsymbol{x}) = P(l_1|\boldsymbol{x}) \prod_{i=2}^{L} P(l_i|l_{i-1}, \ldots, l_1\boldsymbol{x}) \tag{4}$$

This form of joint probability decomposition is known as a probabilistic classifier chain [2]. When learning the conditional probabilities from data, the joint probability becomes dependent on the label order. To remove this dependence, it was suggested that an ensemble of probabilistic classifier chains (EPCC) be used, where each of the classifier chains is constructed using a randomised ordering of labels.

2.3 Probabilistic Label Powerset using Pairwise Coupling

The Label Powerset multi-label classifier has one binary classifier for each label combination, meaning if there are L labels, then at most 2^L binary classifiers are

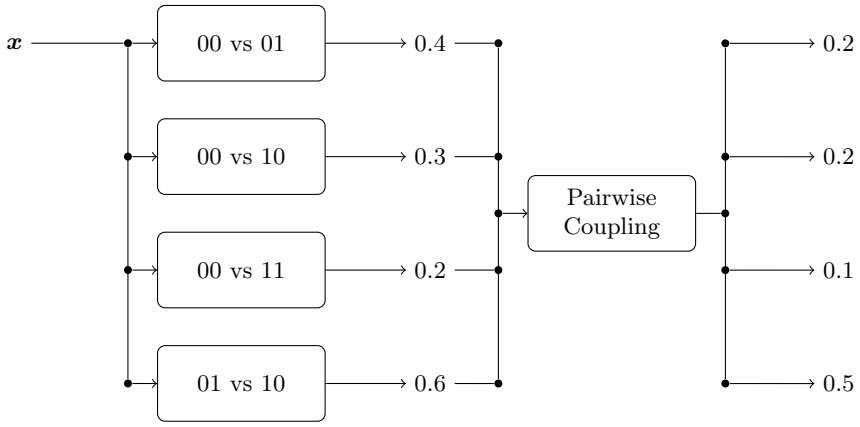

Fig. 1. A Label Powerset multi-label classifier using a set of probabilistic binary classifiers computes the Bernoulli distribution for each label. Pairwise coupling must be used to obtain the categorical distribution over the label set.

required. If we replace the binary classifiers with probabilistic binary classifiers (as done with the previous two methods), we would compute the probability of the given response, independent of all other responses. This means that Label Powerset using probabilistic binary classifiers will provide us with a set of Bernoulli responses rather than a categorical distribution, and hence not provide us with the joint probability. The Bernoulli responses show the probability that the given label set is true and the probability that the given label set false, meaning that the probability of the other label sets are not taken into account. To obtain the joint probability over all label sets, we must compute the probability of a given response, with respect to the probability of all other responses.

To compute the categorical distribution over the powerset of labels, we use multi-class Pairwise Coupling [3] of the powerset of labels. Note that, we can compute the multi-label joint distribution using any probabilistic multi-class method over the 2^L label combinations; we chose Pairwise Coupling because it allows us to construct the probabilistic multi-label classifier using a collection of probabilistic binary classifiers.

The Pairwise Coupling model requires that we train a binary classifier for all *pairs* of label combinations. Each binary classifier allows us to fit a Bernoulli random variable, therefore, for each probabilistic binary classifier, we obtain $p_{i,j}$, the probability of state i being correct and $\bar{p}_{i,j} = 1 - p_{i,j}$, the probability of state j being correct. The set of all pairwise probabilities are then coupled to obtain the complete joint distribution using the following method.

Given each element of a categorical distribution θ_k with $k \in \{1, 2, \ldots, 2^L\}$ states, the probability of being in one state, relative to another is given by:

$$p_{i,j} = \frac{\theta_i}{\theta_i + \theta_j} \tag{5}$$

This gives us at most $\sum_{i=1}^{2^L-1} i = 2^{L-1}(2^L - 1)$ pairwise probabilities to compute the 2^L probabilities of the categorical distribution θ. We find the set of 2^L categorical probabilities θ_k as the categorical distribution that provides the best fit of equation 5 for all k and j using the algorithm from [3]. If the training data contains u unique label combinations, then the joint model will compute only $u(u-1)/2$ pairwise probabilities. Therefore the computation required for the joint model is dependent on the number of unique label combinations available at training time.

Note that each pairwise binary classifier is trained at training time, but the categorical distribution is dependent on the given observations $\boldsymbol{x}$, therefore the coupling is performed during the prediction stage. The pairwise coupling method requires a large number of binary classifiers to perform prediction, but we must remember that each classifier is trained using a subset of the data (only those objects that are associated to the selected pair of label sets for each pairwise classifier), speeding up the training process.

3 Examining the Label Set Distribution

When performing binary classification, it is enough to present the results of the classification as the probability of the class with greatest probability. Once this probability is known, we are able to deduce the probability of the other class. If the probability is close to 0.5, then the classification system has low confidence in its decision. If the probability is close to 1.0, then the classification system has high confidence in its decision.

Multi-label classification consists of many combinations of labels which are all assigned a probability. By reporting only the probability of the label set of greatest probability, we are providing little information to the user. A label set with probability close to 1.0 implies that the system is confident in its prediction, but unlike binary classification, a lower probability has little meaning unless we have the rest of the distribution to compare it to.

The shape of the class distribution gives us a measure of confidence in the predicted results, which is very useful information to any practitioner. For example, let's consider a multi-label problem with only two labels, where the distribution over the four combinations of labels is $\{0.31, 0.3, 0.29, 0.1\}$ for a given value of $\boldsymbol{x}$. If our system predicted the most likely label combination, without providing us with the class distribution, we would accept the result without second thought, which is likely to lead to incorrect predictions. We can see from the distribution that the second and third most probable label sets have similar probability to the label set with the largest probability. This means that there is high uncertainty

in the class of $\boldsymbol{x}$. The label set distribution provided by our system for each sample, allows us to compare the probability of each possible response. This in turn allows us to make a judgement on whether we accept the most likely response as the prediction.

We would expect that a system, very confident in its decision, would provide a label set distribution containing one label set with probability 1 and the remaining label sets with probability 0. A system with no confidence would provide equal probability for all label sets. All other probability combinations would provide varying levels of confidence between these two bounds. If we need to quantitatively measure the confidence level provided by a label set distribution, we can measure the entropy of the distribution:

$$H = -\sum_{\boldsymbol{y}_i \in \mathbb{B}^L} P(\boldsymbol{y}_i) \log\left(P(\boldsymbol{y}_i)\right) \tag{6}$$

which measures the uncertainty provided by the label set distribution [4] ($H = 0$ means no uncertainty), where $0 \times \log(0) = 0$. The range of H depends on the number of label sets in our multi-label problem. To adjust the range to $[0, 1]$, we can use normalised entropy:

$$H^\star = -\sum_{\boldsymbol{y}_i \in \mathbb{B}^L} P(\boldsymbol{y}_i) \frac{\log\left(P(\boldsymbol{y}_i)\right)}{\log\left(2^L\right)} \tag{7}$$

where $H^\star = 1$ is provided when the probability of all class combinations are equal.

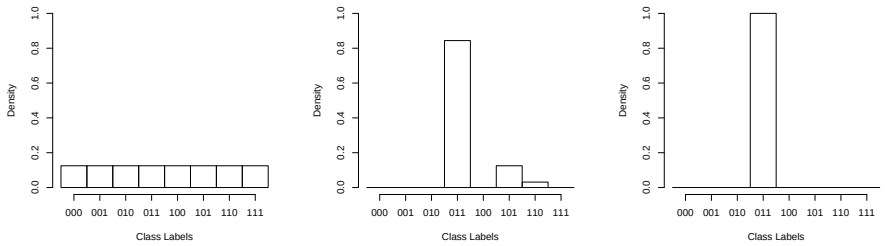

Fig. 2. Multi-label distributions with normalised entropy ($H^\star$) of 1 (left), 0.5 (centre) and 0 (right). In this case there are three labels and hence eight label combinations.

Figure 2 shows us examples of distributions and their entropy (using base e). Note that the only distribution to provide a normalised entropy of 1 assigns all elements with equal probability (as shown in the left plot of Figure 2). A normalised entropy of 0.5 implies that one item has a much greater probability than the others (as in the centre plot of Figure 2). Also, an entropy of 0 implies that

one element has probability 1, with the remaining elements having probability 0 (shown in the right plot of Figure 2).

Note that the measure of entropy is used to identify a systems confidence in its response, but it should not be used to measure the quality of a set of systems without regard to the systems' accuracy. For example a system that provides high levels of entropy for all responses is not worse than a system that provides low levels of entropy. If both systems happen to have a low accuracy, then the former may be preferred over the latter.

4 Using Entropy as a measure of Acceptance

We introduced the topic of measuring uncertainty using the normalised entropy of the multi-label distribution in Section 3. In this section, we will examine if entropy is correlated to accuracy. Our reasoning is:

- If a portion of the sample space can easily be classified, there is little uncertainty in the results and hence the entropy of the multi-label distribution will be low. Low uncertainty also implies that any future predictions should be accurate, meaning that low entropy corresponds to high accuracy.
- If a portion of the sample space is difficult to classify, there will be high uncertainty in the results and hence the entropy of the multi-label distribution will be high. This high uncertainty implies that future predictions are not likely to be correct, meaning that high entropy corresponds to low accuracy.

In this section, we will first describe the data and multi-label models used. We will then examine the relationship of probabilistic multi-label entropy to accuracy.

4.1 Experimental Environment

To perform our investigation, we will use the set of probabilistic multi-label classifiers presented in Section 2: Probabilistic Label Powerset (PLP), Probabilistic Binary Relevance (PBR), and Ensemble of Probabilistic Classifier Chains (EPCC). Each of the methods require the use of a set of binary classifiers that provide a probability measure of its associated label set prediction. In each of our experiments, we use Support Vector Machines with a Radial Basis kernel, where the probabilities were estimated using a Laplace prior [1]. The kernel parameter was kept at the default value of $1/m$, where m is the number of explanatory variables for each observation. The SVM cost parameter for each binary classifier was tuned using 2 shuffle, 5 fold cross-validation on the training data.

The number of probabilistic binary classifiers required for each method is shown in Table 2. We find that the Probabilistic Binary Relevance classifier uses the least number of binary classifiers, while Probabilistic Label Powerset is expected to use the most.

We chose the three data sets shown in Table 3 to perform our analysis; two that are commonly used in multi-label research (Emotions and Scene) and the

Table 2. The number of binary classifiers used by each probabilistic multi-label classifier, where n is the number of binary response variables, u is the number of unique label combinations of response variables within the training data, and e is ensemble size.

Method	Binary classifiers
PBR	n
PLP	$u(u-1)/2$
EPCC	en

Table 3. The data sets used to examine the probabilistic multi-label methods in this article.

Name	Items	Train	Test	Features	Labels	Avg Label Card	Uniq Label Comb
Emotions	593	250	343	72	6	1.8685	27
Scene	2407	1211	1196	294	6	1.0740	15
Stare	373	200	173	44	15	1.3217	42

third from the STARE project[1] (the set of diagnoses from a set of retinal images). The Stare data contains medical diagnoses, where the confidence of prediction is essential and so is a perfect candidate for this research. Note that these data sets are relatively small, but to perform our analysis, we require results for each of the three probabilistic multi-label methods. The Probabilistic Label Powerset method requires $42 \times 41/2 = 861$ binary classifiers for the Stare data set, which consumes most of the CPU time and memory on a modern computer. Therefore it would be difficult to obtain results for larger data sets.

To give perspective on each method, we have presented the training and testing times for each on each data set in Table 4.

To evaluate the classification accuracy of the models, we report results using the 0/1 loss function (if the system returns the correct label set as the most likely label set, it is correct, otherwise it is incorrect). We also examined the use of Hamming and Jaccard similarity for partial matching of labels and found the results to be similar to those reported using the 0/1 loss function.

4.2 Experiment

We computed the normalised entropy for each test sample in each of our three data sets for all three methods. We then examined the accuracy when only considering the k observations from the test set with the lowest response distribution entropy. We expected that when considering the mean accuracy of the chosen k observations, the score should decrease as k increases (since increasing k introduces observations with greater response entropy into the mean calculation).

[1] `http://www.ces.clemson.edu/~ahoover/stare/`

Table 4. The time taken in seconds, for training the model and predicting the state of one object for each method on the Emotions, Stare and Scene data sets.

Methods	Emotions		Stare		Scene	
	Train	Test	Train	Test	Train	Test
PBR	25.51	0.01	21.85	0.96	1147.03	0.02
PLP	101.06	0.66	275.38	3.01	724.75	2.24
EPCC	76.96	0.30	198.38	19.68	3443.75	1.61

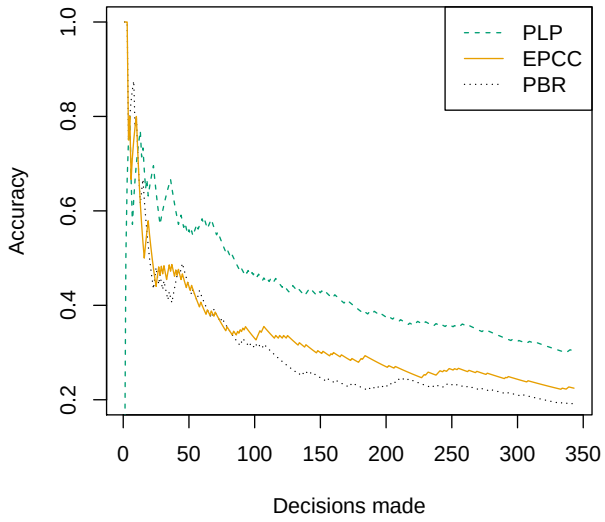

Fig. 3. Using normalised entropy as a measure of prediction acceptance for the Emotion data.

The results of this experiment are shown in Figures 3, 4 and 5 for data sets Emotion, Scene and Stare respectively. Note that k is labelled "Decisions made" in the set of plots, since k can be likened to an entropy threshold in which a practitioner accepts the prediction of the system (makes a decision) when the response distribution is lower than the threshold, while the remainder are discarded as untrustworthy.

The lines in these plots were computed by ordering all of the test observations in order of their normalised entropy. The Accuracy is then computed as the mean of the accuracies of the observation with lowest entropy, the lowest and second lowest, the lowest to third lowest, and so on, until the final value is the mean of all accuracies. Computing the mean in this way causes the first portion of the plot to be jittery since the mean is computed using a small number of samples. As the sample size increases the plot smooths out. The varying number of decisions

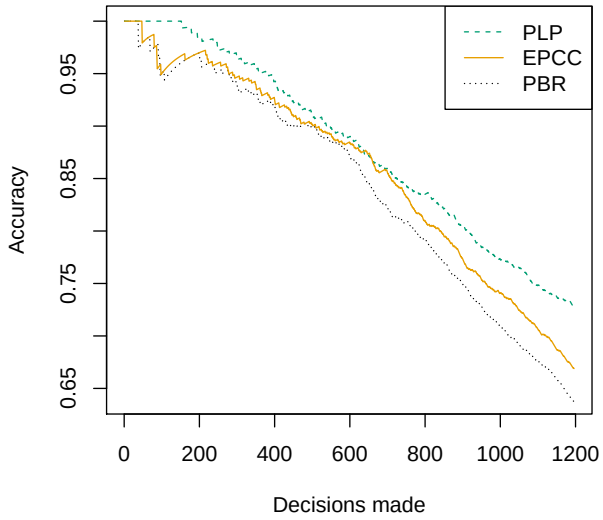

Fig. 4. Using normalised entropy as a measure of prediction acceptance for the Scene data.

made in each of the plots is due to the number of observations available in the associated data test sets.

4.3 Analysis of Results

The plots show the accuracy, ordered by normalised entropy of the response distribution. An increase in the plot line towards an accuracy of 1 means that an accurate prediction was made, a decrease in the line towards an accuracy of 0 means an inaccurate prediction was made. An optimal measure of system confidence would have a plot where the line stays at 1 (placing all of the accurate predictions first), followed by a decrease (placing all of the inaccurate predictions last).

We find that PLP and EPCC have some inaccurate predictions with low normalised entropy (shown by the initial zig-zagging), but then smoothly decrease just after the 50 mark. PBR has the initial zig-zagging but then dips at about the 170 mark, showing a poor ordering of accuracy. The Scene data shows a desired curve from PLP (flat then decreasing), where EPCC and PBR have some initial zig-zagging, but then a decreasing slope. We also find that each of the three methods begin with an initial zig-zag and then gradually decrease for the Stare data.

The general shape of the curves for each plot (flat and then decreasing) show that normalised entropy is a good candidate for measuring confidence of a systems prediction. The variance between methods and plots is due to the different methods used to compute the joint distribution and their behaviour on each data set. We also see from the plots that by accepting only those predictions

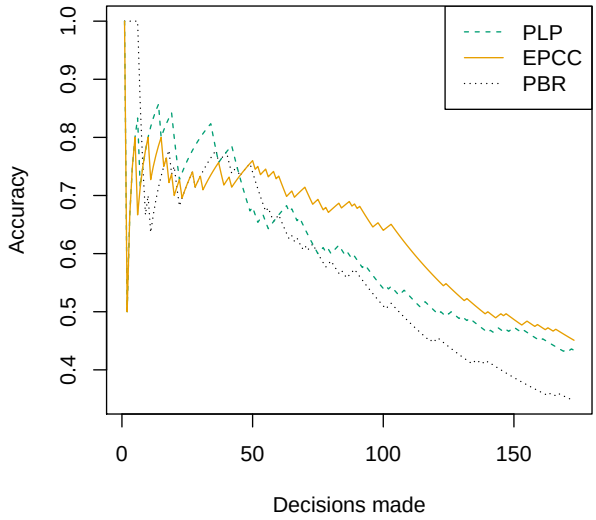

Fig. 5. Using normalised entropy as a measure of prediction acceptance for the Stare data.

that had low normalised entropy, the mean accuracy of accepted predictions is increased by a significant margin for all methods on all data sets. Therefore normalised entropy can be used as a measure of acceptance for probabilistic multi-label classification.

Note that the usefulness of normalised entropy as a measure of acceptance is dependent on the accuracy of the system. If we generated random response distributions, some may have low normalised entropy, but also be inaccurate. Therefore a specialist needs to first choose an appropriate probabilistic multi-label classifier for their data before normalised entropy can be used.

For most of the data sets, there is little difference in accuracy between methods at the low "Decisions made" end of the plots. This is likely to be due to a small sample of points that are simple to classify and hence an independent multi-label classifier is good enough to accurately classify these. The simplicity of their classification would also imply that their multi-label distributions would have low entropy.

5 Conclusion

Multi-label classification allows us to predict the response of many labels at once, using the correlation between the labels to hopefully improve the prediction accuracy. Critical applications of multi-label learning (such as those used in health, military and government) require that predictions are paired with a measure of confidence in the prediction. Probabilistic multi-label classification provides us with a conditional categorical distribution over the powerset of labels, provid-

ing us with an indication of confidence. Unfortunately, it is not always obvious what confidence the system has by observing this distribution. Therefore a single measure of confidence would be useful for specialists using this system.

In this article, we presented a method of determining a probabilistic multi-label system's confidence in its prediction using normalised entropy. We examined correlation of three popular probabilistic multi-label classification method's accuracy with normalised entropy, and found that all three provided a general increase accuracy as the normalised entropy decision threshold reduced. This result shows that we can gauge a systems confidence in its prediction by examining the normalised entropy of the predicted label set distribution.

References

1. Chang, C.C., Lin, C.J.: Libsvm: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST) 2(3), 27 (2011)
2. Dembczyński, K., Waegeman, W., Cheng, W., Hüllermeier, E.: On label dependence and loss minimization in multi-label classification. Machine Learning 88(1-2), 5–45 (2012)
3. Hastie, T., Tibshirani, R.: Classification by pairwise coupling. In: Proceedings of the 1997 conference on Advances in neural information processing systems 10. pp. 507–513. NIPS '97, MIT Press, Cambridge, MA, USA (1998)
4. Shannon, C.E.: A mathematical theory of communication. Bell System Technical Journal 27, 379–423 (July 1948)
5. Zhang, M., Zhou, Z.: A review on multi-label learning algorithms. Knowledge and Data Engineering, IEEE Transactions on 26(8), 1819 – 1837 (2013)