Incorporating Prior Knowledge into Lower Bound of Mutual Information for Feature Selection

Wesley Szamotula
Department of Computer Sciences
University of Wisconsin
Madison, WI 53703
wes.szamotula@gmail.com

Abstract

Feature selection is an essential aspect of machine learning. When used properly it can make classifiers more efficient and more accurate. Filter feature selection is a category of common methods. These are classifier independent methods that use approximations of mutual information to select a subset of features. However, these approximations often rely on poor assumptions that are not usually true. These assumptions can be avoided by using lower bound of mutual information instead. This bound is calculated using an arbitrary measurable Q distribution. The closer this Q distribution is to the data's actual distribution, the better the bound becomes. In this paper I explore a weighted pairwise Q distribution that incorporates prior knowledge of the data's distribution. My experiments show a small improvement over using un-weighted pairwise Q distributions.

1 Introduction

Feature selection is an important aspect of machine learning. Many applications have a large number of features, and some of those features will be irrelevant or redundant. Removing these unnecessary features will improve the speed and accuracy of classifiers. There are several methods of feature selection. Some are classifier dependent, such as wrapper and embedded. Others are classifier independent, such as filter. Filter methods are very useful because they are often much quicker to run than classifier dependent methods and are less prone to over-fitting.

Most filter methods use a measure of mutual information to select a subset of features. However, mutual information is often very difficult to calculate so approximations are used instead. The paper *Variational Information Maximization for Feature Selection* by Gao et al. demonstrated that these approximations are often based on unrealistic assumptions. To estimate mutual information without relying on poor assumptions they formulated a lower bound of mutual information that can be used instead.

The trade-off of using this bound instead of a direct approximation, is that the bound is calculated using an arbitrary measurable Q distribution. The closer the Q distribution is to the actual distribution of the data, the better the bound becomes. In their paper they presented two different Q distributions, the Naive-Bayes Q distribution and the Pairwise Q distribution. In order to obtain good results with this method, the distribution used should be as close to the actual distribution as possible.

To extend upon this method I have explored an additional Q distribution that incorporates prior knowledge to bring it closer to the real distribution, a Weighted Pairwise Q distribution. This distribution uses a weighted geometric mean of conditional distributions over previously selected features. The weights are provided by the user, and can move this distributions closer to the actual distribution of the data. I have validated the usefulness of this method by comparing the feature ranking to the previous method over several synthetic distributions.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

In Sec. 2 I review past research in the area of feature selection. In Sec. 3 I review the derivations for lower bound on mutual information and present the distribution used in this paper. In Sec. 4 I discuss the experimental results of using this new distribution.

2 Past Research

Past research has explored many aspects of filter feature selection. They have covered various topics from derivations of the core principles to applications with specific classification methods.

Using mutual information for selecting features in supervised neural net learning In this paper Battiti noted previous methods of filtering features, like ones based on linear relations, were prone to mistakes because the results were sensitive to scaling in the features. He investigated the application of mutual information as a way to select features since it is independent of the coordinates chosen. The ideal process is to select a subset of features that minimizes the entropy of the system and maximizes the mutual information. This is very hard to do with real world data though, so he used a greedy algorithm as an approximation. For each new feature considered the relevancy (mutual information of the feature and the label) and the redundancy (mutual information of the feature and previous features) were measured. The feature that maximized the relevancy and minimized the redundancy was selected. They validated the improvements using this method over several different experimental data sets.

Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy In their paper Peng et al. derived an equivalent measure of mutual information where relevance between new features and the labels are maximized and the redundancy between new features and existing features are reduced. This new formula is proven to be equivalent to the measure of max-dependency when used in forward feature selection to add one feature at a time. It is also much easier to implement and performs much faster. This measure is calculated by optimizing the following condition in their algorithm:

$$\max_{x_j \in X - S_{m-1}} \left[I(x_j; c) - \frac{1}{m-1} \sum_{x_i \in S_{m-1}} I(x_j; x_i) \right]$$

They also extended the usefulness of this measure by creating a two stage algorithm where filter feature selection is combined with a wrapper method. The algorithm first selects candidate feature sets using max relevance and min redundancy (mRMR). These candidate feature sets are then run through wrapper feature selection to find the optimal feature set. By combining these two methods they are able to take advantage of the speed of filter feature selection and the accuracy of wrapper feature selection. They demonstrated improved results for several classifiers using the mRMR method over just max relevancy.

Conditional Likelihood Maximization: A Unifying Framework for Information Theoretic Feature Selection In this paper Brown et al. focused on deriving a measure of mutual information from a clearly defined object function, the conditional likelihood of the training labels. By approximating the probability distribution of the data with a model q conditioned on the features selected and the parameters, τ, used to predict the labels, they derive the following approximation of the log-likelihood. In the equation below x_θ is the vector of selected features and x is the full feature vector projected onto the selected features.

$$-l \approx E_{xy}\{log \frac{p(y|x_\theta)}{q(y|x_\theta, \tau)}\} + E_{xy}\{log \frac{p(y|x)}{p(y|x_\theta)}\} - E_{xy}\{log\, p(y|x)\}$$

The three terms of this function have direct relationships to feature selection adopted through previous measures. The first term is a measure on how well the q distribution approximates the p distribution. The second term is the conditional mutual information between the class label and the remaining features. The third term is the conditional entropy H(Y|X). By reverse engineering a mutual information selection measure from a clear objective function they have helped validate the legitimacy of past measures. They also conducted experiments using several popular feature selection methods and found that the balance between relevancy and redundancy terms are very important in practice.

Fast Binary Feature Selection with Conditional Mutual Information Flueret's research expanded upon existing measures of mutual information that maximized relevancy and minimized redundancy. By noting a couple important behaviors of these measures he was able to design a much faster algorithm that makes feature selection more useful in very high dimension spaces. Features that have a low score are very likely to remain low as the process goes on. Additionally, scores for features only go down so they don't need to be recalculated on every iteration. With these notes the algorithm for feature selection can reduce the number of times mutual information needs to be recalculated for the next round of feature selection. Specifically, these scores only need to updated if the best one found in an iteration is not better than one from the last iteration. When they compared their algorithm to existing algorithms they were able to demonstrate comparable error rates, while running at a much higher speed.

Variational Information Maximization for Feature Selection In this paper Gao et al. propose an alternate way to measure mutual information for feature selection. The motivation stems from the fact that most mutual information approximations rely on the following assumptions. Features are independent and that features are independent given the class label. These two assumptions can only both be true for very specific, and unrealistic, structures. They are able to avoid utilizing these bad assumptions by instead using the lower bound of mutual information. This bound utilizes an arbitrary measurable q distribution as an approximation of the p distributions. The difference between the bound and the actual distribution is the KL-divergence of the q and p distributions.

They compared the results to several well known feature selection filter methods by generating feature subsets for well-known data sets and comparing cross-validation error for kNN and Linear SVM algorithms. Their feature subsets generated from both naive bayes and pairwise q distributions were able to outperform other methods.

3 Method

In *Variational Information Maximization for Feature Selection* by Gao et al. the lower bound of mutual information is derived from the non-negativity of KL-divergence. They arrive at the following formula where x_s is the feature vector for the features chosen so far and q is an arbitrary measurable distribution:

$$I(x_s; y) \geq \langle ln(\frac{q(x_s|y)}{q(x_s)})\rangle \approx \frac{1}{N} \sum_{x^{(k)}, y^{(k)}} ln \frac{\hat{q}(x_s^{(k)}|y^{(k)})}{\hat{q}(x_s^{(k)})}$$

The validated their results with two different q distributions. A naive bayes distribution and a pairwise distribution that took the geometric mean of conditional distributions. To expand upon their work I have extended the algorithm to run with a weighted geometric mean of conditional distributions using the following formula:

$$q(x_S|y) = q(x_{f_t}|y) \prod_{t=2}^{T} q(x_{f_t}|x_{f<t}, y) \quad q(x_{f_t}|x_{f<t}, y) = (\prod_{i=1}^{t-1} p(x_{f_t}|x_{f_i}, y)^{w_i})^{1/\sum w_i}$$

If the weights are all set to 1 it will be equivalent to the geometric mean validated in the previous work. If some of the users prior knowledge can be used to set weights appropriately it can bring the q distribution closer to the actual distribution of the data. Since this bound was derived from the KL-divergence, the closer the q distribution approaches the actual distribution the better results it will receive.

4 Experiments

To validate the use of the weighted geometric mean I created several synthetic data sets and compared the results to the unweighted geometric mean. The results are presented below by showing how this new method ranks the features and how the old method ranks the features for varying numbers of samples. The features are ranked from best to worst in the tables below.

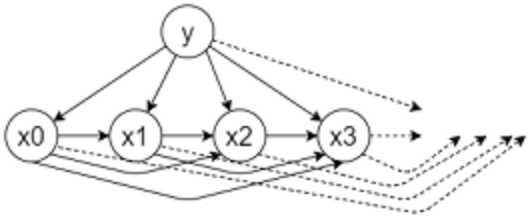

Figure 1: Synthetic distribution where each feature depends on the label and all prior features.

Table 1: Results for synthetic distribution one with weights $\{9, 8, 7, 6, 5, 4, 3, 2, 1\}$

1k samples	
Q Distribution	Ranked features (best to worst)
Pairwise	$x_0, x_6, x_8, x_7, x_5, x_4, x_1, x_3, x_2$
Weighted Pairwise	$x_0, x_1, x_6, x_5, x_8, x_7, x_4, x_2, x_3$
10k samples	
Q Distribution	Ranked features (best to worst)
Pairwise	$x_0, x_1, x_2, x_3, x_8, x_4, x_5, x_6, x_7$
Weighted Pairwise	$x_0, x_1, x_2, x_3, x_4, x_8, x_5, x_6, x_7$

For the first synthetic distribution the weighted geometric mean was able to improve feature selection. When tested on 1,000 samples it moved features x_1, x_2 and x_5 all up to higher positions. When tested on a set of 10,000 it moved feature x_4 to a higher position.

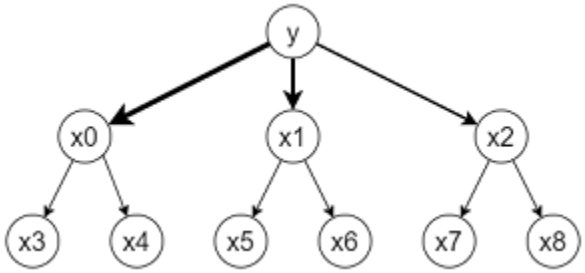

Figure 2: Synthetic distribution with a binary tree structure, where feature x_0 has strongest correlation with the label.

Table 2: Results for synthetic distribution two with weights $\{8, 4, 2, 1, 1, 1, 1, 1, 1\}$

1k samples	
Q Distribution	Ranked features (best to worst)
Pairwise	$x_0, x_1, x_2, x_7, x_6, x_5, x_8, x_3, x_4$
Weighted Pairwise	$x_0, x_1, x_2, x_7, x_5, x_6, x_3, x_8, x_4$
10k samples	
Q Distribution	Ranked features (best to worst)
Pairwise	$x_0, x_1, x_3, x_5, x_4, x_6, x_2, x_7, x_8$
Weighted Pairwise	$x_0, x_1, x_3, x_5, x_4, x_2, x_6, x_7, x_8$

For the second synthetic distribution the weighted geometric mean was able to make a small improvement. For 10,000 samples the feature x_2 was moved to a slightly higher position.

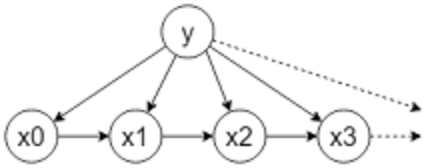

Figure 3: Synthetic distribution where each feature depends on the label and the previous feature.

Table 3: Results for synthetic distribution three with weights $\{3.0, 2.4, 2.2, 2.0, 1.8, 1.6, 1.4, 1.2, 1.0\}$

1k samples	
Q Distribution	Ranked features (best to worst)
Pairwise	$x_0, x_3, x_6, x_8, x_2, x_4, x_7, x_1, x_5$
Weighted Pairwise	$x_0, x_3, x_6, x_8, x_4, x_2, x_7, x_1, x_5$
10k samples	
Q Distribution	Ranked features (best to worst)
Pairwise	$x_0, x_8, x_4, x_2, x_7, x_5, x_1, x_6, x_3$
Weighted Pairwise	$x_0, x_4, x_2, x_6, x_8, x_5, x_1, x_7, x_3$

For the third distribution the weighted geometric mean performed slightly better overall. For 1,000 samples it moved x_2 to a lower position, but for 10,000 samples it moved x_2, x_4, and x_6 to higher positions.

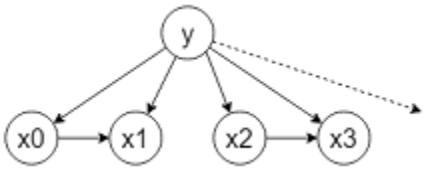

Figure 4: Synthetic distribution where each feature depends on the label, and there are pairs of dependent features.

Table 4: Results for synthetic distribution four with weights $\{2, 1, 2, 1, 2, 1, 2, 1, 2\}$

1k samples	
Q Distribution	Ranked features (best to worst)
Pairwise	$x_4, x_2, x_8, x_0, x_6, x_3, x_5, x_1, x_7$
Weighted Pairwise	$x_4, x_2, x_8, x_0, x_6, x_3, x_5, x_1, x_7$
10k samples	
Q Distribution	Ranked features (best to worst)
Pairwise	$x_0, x_2, x_8, x_6, x_5, x_3, x_4, x_7, x_1$
Weighted Pairwise	$x_0, x_2, x_8, x_6, x_4, x_5, x_3, x_7, x_1$

For the fourth distribution the weighted geometric mean was able to make a slight improvement. For 10,000 samples the feature x_4 was moved to a higher position.

Overall the weighted geometric mean was able to make a small improvement on the results of the unweighted geometric mean. Strong improvements were seen for the first distribution, and small improvements were seen for the other three.

5 Future Directions

There are two main directions I would take this project next. The first would be improving the evaluation of the changes. With some additional data sets where the mutual information is known, I would be able to measure the distance between the lower bound and the ground truth. This would give a concrete example of how the q distribution is able to approach the actual p distribution.

The second focus would be on extending the tests to additional q distributions. Some other common measurable distributions would be a useful addition so users could select the one that best matches their data. Additionally, this algorithm was designed for use with forward feature selection so the q distributions are calculated with respect to previously selected features. It could be modified to support reverse feature selection as well. This would give more options for how the features subset could be identified.

6 Conclusion

Feature selection is a very important aspect of machine learning. Selecting a representative subset of features can improve both the speed and accuracy of classifiers. Most previous methods of filter feature selection use approximations of mutual information to select high value features. However, these approximations often rely on some bad assumptions. By instead using a lower bound of mutual information we can make approximations without these assumptions. The lower bound is calculated using an arbitrary measurable q distribution, and the closer this distribution is to the actual distribution the better the bound becomes.

Past research explored a naive bayes and pairwise q distribution and found improvement over most previous methods. In this paper we have extended their work to a weighted pairwise distribution. This allows us to incorporate prior knowledge about the distribution to better approximate it. Over several synthetic distributions this new method is shown to better rank the available features. This gives us a small improvement over the unweighted pairwise q distribution.

References

[1] Battiti, R (1994) Using Mutual Information for Selecting Features in Supervised Neural Net Learning. *IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 5, NO. 4* pp. 537-550

[2] Brown, G. & Pocock, A. & Zhao, M. & Lujan, M. (2012) Conditional Likelihood Maximisation: A Unifying Framework for Information Theoretic Feature Selection, *Journal of Machine Learning Research 13* pp. 27-66

[3] Peng, H. & Long, F. & Ding, C (2005) Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 27, NO. 8* pp. 1226-1238

[4] Fleuret, F (2004) Fast Binary Feature Selection with Conditional Mutual Information. *Journal of Machine Learning Research 5* pp. 1531-1555

[5] Gao, S & Steeg, G & Galstyan, A (2016) *Variational Information Maximization for Feature Selection* 30th Conference on Neural Information Processing Systems