Semi-Supervised Feature Selection with Constraint Sets

Prajakta Kulkarni
ME Candidate
K. K. W. I. E. E. R., Nashik

S. M. Kamalapur, PhD
Assistant Professor
K. K. W. I. E. E. R., Nashik

ABSTRACT
In machine learning classification and recognition are crucial tasks. Any object is recognized with the help of features associated with it. Among many features only some leads to classify object correctly. Feature selection is useful technique to detect such specific features. Feature selection is a process of selecting subset of features to reduce number of features (dimensionality reduction). Semi-supervised feature selection is difficult due to scarcity of labeled samples. Here constraint based approach is proposed to efficiently select features from semi-supervised data. Constraint based approach is selected as it incorporates supervised information in processing. In the absence of labels, features can be evaluated based on locality preserving ability. Hence for semi-supervised data, properties of both labeled and unlabeled data are combined tochoose good features. Constraint based Laplacian score is used to find weight of features. To eliminate redundant features mutual information is calculated and graph based method is used to remove redundant features. Classification accuracy for different dataset is measured to check performance of system.

General Terms
Pattern Recognition, Machine Learning

Keywords
Constraints, feature selection, redundant, relevant, semi-supervised

1. INTRODUCTION
In machine learning, task of classification and recognition is automated. Machine gets learned from different examples and their associated features. But when data is having large number of features like thousand or multiple of thousand then it creates problems. For example areas which processes on images or biological data, categorizes texts, detects the unusual patterns, etc. faces this situation. More number of features increases the space and time complexity of learning. This problem is solved by dimensionality reduction technique. Using this technique only useful dimensions or features are retained and remaining are discarded. Feature extraction and feature selection are two ways of dimensionality reduction. Feature extraction is a process which generates asmall new set of features by combining existing oneswhereas feature selection selects the subset of featuresfrom existing ones. Feature selection is widely usedas it does not alter the original features.Feature selection aims to find a subset of featureswhich will give more information about data than anyother combination of features. Feature selection reducesthe computational complexity and increases predictionaccuracy. Let *F* be the number of available features then2^F different subsets of features are possible. The taskof feature selection is to select a subset with optimumnumber of features so thatfurther processing can befocused on them. Feature selectioncan be applied to different types of trainingdata i.e. supervised, unsupervised and semi-supervised.In supervised feature selection classes(or labels) arepredefined whereas for unsupervised feature selectionclasses are not predefined. For semi-supervised datasome samples are having predefined classesand remaining has no predefined class. Many methods for supervised and unsupervised feature selection are present in literature. But these methods are not suitable for semi-supervised samples. Hence proposed system is dealing with efficient semi-supervised feature selection.In recent years useof pairwise constraints is increasing to improve the performance of semi-supervised feature selection. Constraintsrepresent the background knowledge about data andhence guides the process of feature selection. With thehelp of constraints the available labeled information canbe efficiently used to select relevant and non-redundantfeatures. Relevant feature is one which gives useful informationabout decision feature i.e. class and redundantfeature is one which does not provide more informationthan the subset currently selected.

Here are the notations used to represent the semi-supervised feature selection. Let $Y = \{ y_1, y_2, \dots , y_N \}$ be the set of N semi-supervised samples. Hence$Y_l = \{ y_1, y_2, \dots , y_l \}$ and $Y_u = \{ y_{l+1}, y_{l+2}, \dots , y_{l+u} \}$ are the two subsets of Y indicating labeled samples and unlabeled samples respectively where$N = l + u$. Let $A_1, A_2, \dots, A_F$ denote features or attributes of Y and $\{a_1, a_2, \dots , a_F\}$ be corresponding feature vectors that record feature value for every sample.

In next section brief literature of feature selection is described. Section III describes proposed approach along with its block diagram. The evaluation of system is discussed in section IV and section V is conclusion.

2. RELATED WORK
Features can be discarded without decrease in performance of learning. Selection of features is dependent on measureused to evaluate features. In supervised feature selection, labels are presentfor all samples. Therefore, evaluation measure or functionmakes use of labels to find the relevant features. Theconventional methods like Relief[2], Fisher Score[3], FastCorrelation Based Filter(FCBF)[4] evaluates the featuresby finding correlation with labels.For unsupervised instances there are no labels whichcan help to search relevant features. In this situationthe feature is evaluated by its ability to maintain certainproperties of data like variance or separability Variance score,Laplacian score[5], SPEC[6] uses this evaluation for unsupervisedfeature selection.

For unsupervised training data if supervision information is available in some form then it is more useful to select good features. Pairwise constraints [7] are used efficiently in clustering process. Constraints states whether two samples can come together or not. They provide the guidance to form the proper clusters. In clustering process pairwise constraints are provided by users. Similar kind of information can be provided by labels or classes.

In semi-supervised feature selection the samples which are having labels will help to form such constraints [8]. Now use of only labeled samples is not sufficient as they are very less as compared to unlabeled samples. Hence unlabeled samples should also be combined with them. Semi-

superviseddimensionality reduction(SSDR)[9], sSelect[11], Constrained Laplacian score(CLS)[12], Constraintselection based feature selection(CSFS)[13] approachesmakes use of both labeled and unlabeled informationfor semi-supervised feature selection.

The above mentioned methods use labeled information to form constraints. These constraintsare presented in two forms: must-link and cannot-linkset. The must-link set contains pairs of samples whichcan come together and cannot-link set contains that pairs which cannot come together. While evaluatingthe features, conditional information provided by constraintsis incorporated. So, such features get selected whichpreserves the provided constraints.

Z. Zaho and H. Liu addressed the problem of semisupervisedfeature selection and presented a solutioncalled sSelect. This algorithm assumes that "if points arein the same cluster, they are likely to be of the sameclass". It constructs the cluster indicator from featurevector and evaluates feature relevance using labeledand unlabeled data. But the limitation of algorithm isthat it can handle only two class problem. Usingpairwise constraints D. Zang et al. given an algorithmnamed SSDR. It evaluates the features based on how itpreserves the intrinsic structure of data andconstraintsprovided. But it cannot preserve the local structure ofdata. Laplacian score is best suited method for unsupervisedfeature selection as it can preserve the localgeometric structure of data. Hence in CLS which is semi-supervisedfeature selection approach, Laplacian scoreis integrated with pairwise constraints. This helps infinding more relevant features. I. Davidson et al.[14] hasshown that all pairwise constraints are not useful andhence provided a way to measure the utility of constraintsets. These measures are applied in CSFS and it hasshown improvement in selecting features.All these methods are rank-based methods. Due to this features which are correlated with each other gets similar rank. So, it may happen that both features get selected in final subset which is not required. L. Yu and H. Liu[15]has addressed this problem and suggested a framework to eliminate redundancy based on correlation measure. A graph based approach is presented in CSFSR [16]. In CSFSR mutual information is used to know the relation between features and highly related features are iteratively removed.Removing the redundancy from subset showedthe increase in performance of learning.

3. PROPOSED SYSTEM

Semi-supervised feature selection imposes a challenge as both labeled and unlabeled samples are present together. As the data is sampled from same population they describe thesame objective concept. To understand this concept use ofboth partitions of data is necessary. Constraint set basedfeature selection is popular while dealing with semi-supervised data because it is the only way to incorporate labeled information. Feature selection is achieved through next main steps. First labeled constraint sets are formed using the supervised samples. Then features are evaluated using constrained Laplacian score. This evaluation function is chosen as it makes use of both labeled and unlabeled data together. This will give the subset of relevant features. To eliminate redundancy, if any, from subset of features graph based approach is used which iteratively removes [16] the redundant features. Semi-supervised data has large number of unlabeled samples. To reduce complexity of feature selection, unlabeled data is clustered. Proposed system architecture is as shown in figure 1.

3.1 Details of System

3.1.1 Form constraint sets

Labels or classes provides information about samples. In feature selection labels are used to find correlation between feature and label. This determines the relevancy of feature. Now here pairwise constrains are formed and selected using labeled data. Two constraint sets are formed i.e. *ML* and *CL*as shown below.

$$ML = \{(y_i, y_j), \ldots\} y_i, y_j \in Y_L \text{ and has same label}$$

$$CL = \{(y_l, y_m), \ldots\} y_l, y_m \in Y_L \text{ and has different label}$$

Davidsonetal. [17]showed that all constraints are not participating in acquiring good accuracy,but some constraints decrease accuracy. To know which constraint sets are useful,coherence between constraints is measured. Incoherent constraint sets are removed and now selected constraint sets are denoted by *ML'* and *CL'*.

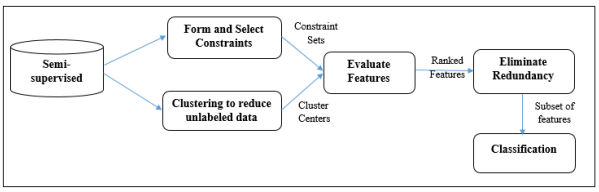

Fig 1: System Architecture

3.1.2 Cluster unlabeled data

Clustering of data is applied to reduce number of samples. Large unlabeled data increases the processing time. To minimize this time clustering is carried out. Cluster centers works as representative of respective cluster. Hence only cluster centers are included further for selection of features. While forming clusters it is checked whether samples having same labels are in same cluster or not [7]. Hence supervision information helps to form proper clusters. Unlabeled data is replaced by cluster centroids. Number of clusters are decided by distribution of data. If number of classes is known then that many clusters are formed. Cluster centers and labeled samples are together used further for evaluating features.

3.1.3 Evaluation of features

After preprocessing unlabeled data now there is time to evaluate features. Rank based feature selection approach is efficientfor high dimensions as it scales well. Properties ofunlabeled data can be revealed by nearest neighborgraph structure. The data which has same concept orclass will be nearer to each other. Laplacian score[5] forfeature selection is based on same concept. Features areevaluated according to variance and locality preservingability. Labeled set information is integrated with nearest neighbor graph. The score φ_r of r^{th} feature isformulated as,

$$\varphi_r = \frac{\sum_{i,j} (a_{ri} - a_{rj})^2 (S_{ij} + N_{ij})}{\sum_{i,j} (a_{ri} - \alpha_{rj}^i)^2 D_{ii}}$$

where,

$$S_{ij} = \begin{cases} corr(y_i, y_j) & if\ y_i\ and\ y_j\ are\ neighbors \\ 0 & otherwise \end{cases}$$

and,

$$\alpha_{rj}^i = \begin{cases} a_{rj} & if\ (y_i, y_j) \in ML' \\ \mu_r & if\ y_i \in Y_U \\ a_{ri} & otherwise \end{cases}$$

and,

$$N_{ij} = \begin{cases} -corr(y_i, y_j) & if\ y_i\ and\ y_j\ are\ neighbors\ and\ (y_i, y_j) \in ML'\ or \\ & y_i\ and\ y_j\ are\ not\ neighbors\ and\ (y_i, y_j) \in CL' \\ \left(corr(y_i, y_j)\right)^2 & if\ y_i\ and\ y_j\ are\ neighbors\ and\ (y_i, y_j) \in CL'\ or \\ & y_i\ and\ y_j\ are\ not\ neighbors\ and\ (y_i, y_j) \in ML' \\ 0 & otherwise \end{cases}$$

and,

$$corr(y_i, y_j) = \frac{\sum_m (y_{im} - \bar{y}_m)(y_{jm} - \bar{y}_m)}{\sqrt{\sum_m (y_{im} - \bar{y}_m)^2} \sqrt{\sum_m (y_{jm} - \bar{y}_m)^2}}$$

The term a_{ri} is value of r^{th} feature fori^{th} sample. HereS_{ij} [5] is weight matrix obtained using nearest neighborgraph and its value is nonzero i.e. correlation betweeny_i and y_j if y_i among the nearest neighbor of y_j orvice a versa and *D* is diagonal matrix obtained from *S*.The nearest neighbors are find out to know the localgeometric structure of data. In nearest neighbor graph,nodes are samples and edge weight is distance betweenthem. Ideally the two samples from same class hasminimum distance. So basic idea is that, a featureis good, if it has more variance and instances fromsame class are nearer to each other. Mean ofr^{th} featureis denoted as $\mu_r$the N_{ij} term is added to score tointegrate labeled information with Laplacian score. Scoreof feature is increased in two bad cases: one when twoinstances has same label but not neighbors and secondwhen two instances does not have same labels but they are neighbors. According to this function feature should get minimumscore value to get selected as maximum varianceis preferred. Hence features are arranged in ascending orincreasing order. From this first *S* features are selected assubset. The subset obtained is relevant but it may includeredundancy. Redundant features leads to deterioration oflearning performance and increase in time complexity[4].Hence redundancy analysis is carried out on selectedsubset.

3.1.4 Eliminate redundancy

High correlation between features represent the featureredundancy. The two features are redundant if theyare highly correlated with each other. Such featuresshould be removedas they introduce the storage complexityissue without providing any extra information. The measure of mutual information is usedwhich can be directly calculated from correlation coefficient.Maximum spanning tree (MST) is used to eliminatethe redundant features and to keep the strong relevantones. Prims algorithm gives minimum spanning tree. The same concept is applied to obtain MST. Following stepsare to be followed to obtain non-redundant subset. Illustrationof steps is also shown but only four features areconsidered to serve the purpose of illustration.

- Find mutual information between every pair of features A_r and A_c

$$I(A_r, A_c) = -\frac{1}{2}\log(1 - \rho^2(A_r, A_c))$$

Mutual information *I* expresses dependence between joint distributions of both features. ρ is correlation coefficient between two features.

- Now construct the complete and weighted graph*G(V,E)*. *V* is set of relevant features (vertices) and *E* is set if edges.
- Apply Prims based MST algorithm [18] to obtain *G'(V', E')*. *E'* contains edges with maximum weight. Arrange feature according to their relevancy and iteratively remove redundant features.
- The set of remaining features is relevant and non-redundant.

4. EVALUATION OF SYSTEM

Datasets used for feature selection are high dimensional. All instances in datasets arelabeled. Names of datasets and other information is listed in Table 1. PIE and ORLare standard face image datasets and TOX and CLL are microarray datasets. Labelsare present for all data. System presented here is for semi-supervised feature selection.Hence for experimental purpose 10% of total samples are randomly selected as labeledand rest are considered as unlabeled. No other preprocessing is applied on datasets.

Experiments on proposed system are carried out using four above mentioned datasets.Aim of experiment is to check effectiveness of selected features. Results obtained arecompared with other similar systems of feature selection.If selected features are able to classify data correctly then selected features are efficient.Classifier is required to classify objects. Here Support Vector Machine(SVM) isemployed for classification purpose. Using 3-fold cross validation parameters of SVM are tuned. LIBSVM library is used to implement SVM.Data is divided into training and testing by stratified sampling. 50% of data is selectedfor training and remaining data is for testing. Classification accuracy is obtainedon testing data.

4.1 Select Top 100 Features

By selecting top 100 features from relevant and non-redundant subset accuracy is obtained.20 runs are taken for each dataset and averaged accuracy is considered. Selectionof features is affected by distribution of data among class and number of data samplesavailable.The accuracyobtained for 'PIE' dataset is maximum as compared to all otherdatasets as shown in figure 1.

Table 1: Dataset

Dataset Name	Instances	Dimensions	Classes
PIE	210	2420	10
ORL	400	1024	10
CLL	111	11340	3
TOX	171	5748	4

Pattern obtained by 'PIE', 'CLL', and 'TOX' showsthat after selecting specific number of features, there is no change in accuracy.

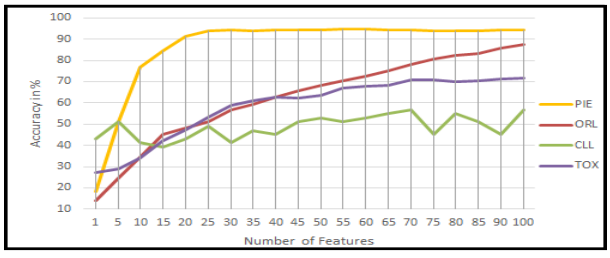

Fig 2: Classification Accuracy across Different Number of Selected Features

4.2 Comparison with Similar Systems

Proposed system is compared with other similar systems as shown in table 2. In table 2, number of selected features are shown in round brackets. Observations suggeststhat for PIE and TOX datasets accuracy obtained is better than othermethods. Proposed system showed improved accuracy with less or equal number of features. So,compromise between small label information and geometric structure of data is usefulfor semi-supervised feature selection.

Table 2: Comparison of Classification Accuracy (in %) with Similar Systems

Dataset	sSelect	C4	CLS	Proposed System
PIE	84.33(88)	86.8(101)	87.8(88)	94.8(88)
ORL	85.25(82)	96.56(97)	96.76(93)	81.06(88)
CLL	62.03(61)	69.44(66)	64.44(62)	58.02(40)
TOX	60.56(58)	62.9(61)	62.39(48)	62.16(45)

5. CONCLUSION

Feature selection when applied reduces the training time and storage requirements.Conventionally there are two types of feature selection: supervised and unsupervised.Supervised feature selection is superior to unsupervised one. But it is not possibleto obtain labels for all unsupervised data. Hence semi-supervised feature selection isimportant. Efficient semi-supervised feature selection method is proposed here. Supervisedinformation is used to form constraints. Coherent constraints are selected by findingtheir overlap. As data is semi-supervised, local structure of data is used to evaluatefeatures. Selected constraints are incorporated while evaluating features. Relevancy offeatures is decided based on score obtained using constraint based Laplacian function.Features should be relevant as well as non-redundant. Redundancy analysis is carriedout using maximum spanning tree. By iterating through spanning tree features havinghigh mutual information are removed. The proposed system achieves better classificationaccuracy with either same number of features or reduced number of features ascompared with other systems.

6. ACKNOWLEDGMENTS

I am very thankful to my guide Prof. Dr. S. M. Kamalapur for her continuous encouragement and help for fulfilling this work. Also I am thankful to other staff members and friends for their assistance in completing this work.

7. REFERENCES

[1] Zhao Z. and Liu H., C. 2003. Semi-supervised feature selection via spectralanalysis. .

[2] Robnik-Sikonja M. and Kononenko, J. 2003.Theoretical and empiricalanalysis of relief and relief.

[3] Duda R. O., Hart P. E., and Stork D. G., 2000.Pattern Classification.

[4] Yu L. and Liu H., 2004. Efficient feature selection via analysis of relevance and redundancy. J. Mach. Learn. Res. 5 (Oct. 2004), 1205-1224.

[5] He X., Cai D. and Niyogi P., C. 2005. Laplacian score for feature selection.

[6] Zhao Z. and Liu H., I. C. 2007. Spectral feature selection for supervised and unsupervised learning.

[7] Davidson I. and Basu S., 2007. A survey of clustering with instance level constraints. ACM transactions on knowledge discovery from data.

[8] Zhang D., Chen S and Zhou Z., 2008. Constraint score: a new filter method for feature selection with pairwise constraints. Pattern recognition(2008) 1440-1451.

[9] Zhang D., Zhou Z. and Chen S., I. C. 2007. Semi-supervised dimensionality reduction.

[10] Benabdeslem K. and Hindawi M., 2011. Constrained laplacian score for semi-supervised feature selection. Proc. ECML-PKDD, Athens, Greece (2011), 204-218.

[11] Hindawi M., Allab K. and Benabdeslem K., 2011. Constraint selection based semi-supervised feature selection. Proc. IEEE ICDM, Vancouver, BC, Canada (2011), 1080-1085.

[12] Allab K. and Benabdeslem K., 2011. Constraint selection for semi-supervised topological clustering. Proc. ECML-PKDD, Athens, Greece(2011). 28-43.

[13] Davidson I., Wagastsff K. and Basu S., 2006. Measuring constraint set utility for partitional clustering algorithms.

[14] Peng H., Long F. and Ding C., 2005. Feature selection basesed on mutual information: criteria of max-dependency, max-relevance and min-redundancy. IEEETrans. Pattern Anal. Mach. Intell.(Aug. 2005), 1226-1238.

[15] Hindawi M. and Benabdeslem K., 2014. Efficient semi-supervised feature selection: constraint, relevancy and redundancy. IEEE Trans. Knowledge and Data Engg. (May 2014).

[16] Chung F., 1997. Spectral graph theory.

[17] I. Davidson, K. Wagstaff, and S. Basu, 2006. Measuring constraint set utility for partitional clustering algorithms, in Proc. ECML/PKDD.

[18] Cormen T. H., Stein C., Rivest R. L. and Leiserson C. E., 2001. Introduction to algorithms.