Discriminative Clustering Based Feature Selection and Nonparametric Bayes Error Minimization and Support Vector Machines (SVMs)

K. Saranya[1], T. Deepa[2]

[1]Research Scholar, Department of Computer Science, Sri Ramakrishna College of Arts and Science for Women, Bharathiar University, Coimbatore, Tamilnadu, India

[2]Assistant Professor, Department of Computer Science, Sri Ramakrishna College of Arts and Science for Women, Bharathiar University, Coimbatore, Tamilnadu, India

Abstract: ***In recent years feature selection is an eminent task in knowledge discovery database (KDD) that selects appropriate features from massive amount of high-dimensional data. In an attempt to establish theoretical justification for feature selection algorithms, this work presents a theoretical optimal criterion, specifically, the discriminative optimal criterion (DoC) for feature selection. Computationally DoC is tractable for practical tasks that propose an algorithmic outline, which selects a subset of features by minimizing the Bayes error rate approximate by a non-parametric estimator. A set of existing algorithms as well as new ones can be derived naturally from this framework. In the proposed Discriminative Clustering based feature Selection algorithm (DCBFS) minimum spanning tree is constructed to group the similar feature from the dataset. Also, efficient algorithms for multiple kernel learning and best feature selection algorithm are introduced. Kernel function called Gaussian Radial basis Polynomial Function (GRPF) is introduced in order to improve the classification accuracy of Support Vector Machines (SVMs) for both linear and non-linear data sets. The aim is Support Vector Machines (SVMs) with different kernels compared with back-propagation learning algorithm in classification task. Finally the proposed algorithm is improved in terms of accuracy and time compared to the existing algorithm.***

Keywords: Feature selection, accuracy, classification, discriminative optimal criterion.

1. Introduction

A "feature" or "attribute" or "variable" refers to an aspect of the data [1]. The Data mining techniques are embedded in Feature subset selection by eliminating the irrelevant features from the dataset. Since Irrelevant features might consume negative effects on a prediction task. Furthermore, the computational complication of a classification procedure might suffer after the curse of dimensionality affected by numerous features. When a dataset contains numerous irrelevant feature variables and simply a few examples, it results in over fitting. In addition the records or data are typical, categorized by fewer variables. Feature selection has been applied in many fields such as multimedia, image classification and biometric recognition. Feature selection methods can be divided into;

- Filter approach
- Wrapper approach
- Embedded approach

The filter approach computes the feature evaluation weight but without performing classification of data, eventually finding the 'good' subset of features. The wrapper-based methods employ some inductive classification algorithms to evaluate the goodness of subset of features being selected.

The Embedded approaches is a Specific learning machine that performs variable selection (implicitly) in the process of training E.g. WINNOW-algorithm (linear unit with multiplicative updates) Feature selection can significantly increase the performance of a learning algorithm in terms of both accuracy and computational time but it is not easy in the existing methods.

Optimal criterion for feature selection, namely the discriminative optimal criterion (DoC), is used as a complementarity to the representative one (referred to as representative optimal criterion (RoC)). The DoC directly attempts to maximize the classification accuracy and naturally reflects the Bayes error in the objective. Compared to RoC, DoC is practically positive in supervised classification. However, DoC is computationally intractable as it involves unknown probabilistic densities. To make DoC practical, an algorithmic framework for feature selection has to be proposed, which selects a subset of features by minimizing the Bayes error estimated by a nonparametric estimator.

2. Related Works

Wrapper or Embedded methods [2] [3] use a specific type of classifier to evaluate the quality of a feature subset of high dimensional data and select the optimal feature subset by minimizing the training error at chosen classifier. The wrapper approaches of feature selection aim to find the minimum discriminative features to reach the high classification accuracy, while the filer approaches are derived to compute the 'best' subset of features in terms of some criteria. However, the inherent nature among features such as function regulation and frequent patterns has been ignored in both filter and wrapper approaches. The major disadvantage of those methods is that each subset of features is evaluated regarding their dependencies, thereby ignoring the functional regulation among the features [4].

Kenneth et al. [7] proposed a technique for non-parametric Evaluation of Renyi's Entropy to train the preprocessor by maximizing the mutual information between the class labels

and the output of the preprocessor. The results also directly associated with the Bayes error for classification [7].

Relief algorithm [5] recently interpreted as a method that optimizes the average heuristically margin [6], [7] [8] though the secret behind the meaning of the boundary is unclear. To the deep understanding of existing feature selection methods and strategies for the expansion of novel algorithms, it is extremely necessary to find optimal feature selection evaluation criteria that have sound theoretic sympathetic condition.

(Buturovi Hc) Proposed to use the k-NN estimate of Bayes error in the transformed space as an optimization criterion.Here, the Bayes error is approximated by upper and lower error bounds for the appropriate range of k, and then minimized using the simplex algorithm in the space spanned by transformation matrix coefficients. Apparently, the performance of this method depends on k.

An advantage of this approach is that the number of features necessary for classification without serious information loss can be predicted.

(Hild, et al., 2006) [9] Proposed the Feature extraction method by utilizing an error estimation equation based on the Bhattacharyya distance. A new criterion for feature extraction was proposed to use classification errors in the transformed feature space, which are predictable using the error estimation equation.

N. Vasconcelos (2003) [10] proposed Main results of a theoretical characterization of the problems for which the principle is guaranteed to be optimal in the infomax sense.

Kai Yu et al.,(2003) proposed the Collaborative filtering has been very successful in both research and practice. However, important research issues remain to be addressed in order to overcome two fundamental challenges in collaborative filtering. (1)Scalability: Existing collaborative filtering algorithms can deal with thousands of consumers in a reasonable amount of time, but modern E-Commerce systems need to handle millions of consumers efficiently; (2)Accuracy: Consumers need recommendations they can trust to help them find products they will like.

3. Proposed Methodology

In this chapter the proposed concept, the discriminative feature selection and discriminative feature selection result with Nonparametric Bayes Error Minimization and SVM based classification algorithm. The proposed methodology discusses about the feature selection using DCBFS algorithm;

- Discriminative Clustering Based Feature Selection with SVM
- Discriminative Clustering Based Feature Selection with GRPF-SVM

Existing both SVM and KNN algorithm doesn't select best features, clustering based feature subset selection will differs from normal feature selection algorithm. Clustering based feature selection algorithm group the similar features in the dataset.

3.1 Discriminative clustering based feature selection

The irrelevant feature removal is straightforward once the right relevance measure is defined or selected, while the redundant feature removal is a bit of sophisticated. In our proposed algorithm Discriminative clustering based feature selection, it involves;

- The structure of the minimum spanning tree from a weighted complete graph;
- The partitioning of the MST (minimum spanning tree) into a forest with each tree representing a cluster;
- The group of representative features from the clusters.

Discriminative clustering based feature selection algorithm logically consists of three steps:

1) Removing irrelevant features,
2) Constructing an mst from relative ones, and
3) Partitioning the mst and selecting representative features.

Clustering Based Feature Selection Algorithm Input: D($F_1, F_2, \ldots\ldots\ldots F_M, C$) the given dataset θ *the* T-relevance threshold

Output: S-selected feature subset

1. For i=1 to m do
2. T-relevance = SU(F_i, C)
3. If T-relevance$\geq \theta$ then
4. S=S$\cup\{F_i\}$;
5. G=NULL;G is the complete graph
6. For each pair of features $\{(F_i', F_j') \subset S$ do
7. F-correlation=SU(F_i', F_j')
8. Add F_i' *and /or* $,F_j'$ to G with F-correlation as the weight of the corresponding edge
9. Forest =Minspantree
10. For each edge $E_{ij} \in$ *Forest* do

$$(F_i', F_j') < SU(F_i', C) \wedge SU(F_i', F_j') < SU(F_j', C)\ then$$

11.if SU
12. Forest =Forest -E_{ij}
13. S=$\emptyset$
14. For each tree $T_i \in$ *Forest* do
15. F_R^j =argmax $argmax_{F_k' \in T_i}$ SU(F_k', c)
16. $S = S \cup \{F_R^j\}$;
17. Return S

3.2 Discriminative Clustering Based Feature Selection with SVM

From this feature subset selection result in the above algorithm then classify the feature in the data .Support

vector machine based feature subset selection algorithm is performed to classify the data in the feature subset. The basic SVM takes a set of input data as feature subset result from the Discriminative clustering based feature selection and predicts, for each given input, which of two feasible classes forms the output, making it a non-probabilistic binary linear classifier.

3.3 Discriminative Clustering Based Feature Selection with GRPF-SVM

Support vector machine classification is choosing a suitable kernel of SVMs for a particular application, i.e. various applications require different kernels to get dependable classification results. It is well known that the two typical kernel functions often used in SVMs are the radial basis function kernel and polynomial kernel. More recent kernels are presented to handle high dimension data sets and are computationally efficient when handling non-separable data with multi attributes. However, it is not easy to find kernels that are able to achieve high classification accuracy for a diversity of data sets. In order to create kernel functions from existing ones or by using some other simpler kernel functions as building blocks, the closure properties of kernel functions are essential.

Gauss RBF

Combine POLY, RBF, and PRBF into one kernel to become:

$$GRPF(x,z)=\left(\frac{d+r.exp(-\|X-z\|^{r}/(r.\sigma^{2}))}{r+d}\right)^{d+1}$$

$$\theta^{0} = arg \min_{\theta} T(\alpha^{0},\theta)$$

where σ is a statistic distribution of the probability density function of the input data; and the values of r(r >1) and d can be obtained by optimizing the parameters by means of the training data. The proposed kernel has the advantages of generality. However, the proposed Gaussian and polynomials kernel function by setting d and r in different values. For example if d =0, Exponential Radial when r= 1 and Gaussian Radial for r= 2 and so on. Moreover different kernels can be obtained by optimizing the parameters using the training data. GRPF depends on two parameters d and r, encoded into a Vector θ = (d, r). Thus consider a class of decision functions parameterized by α, b, θ:

$$f_{\alpha,b,\theta}(x) = sign(\sum_{i=1}^{l} \alpha_i y_i GPRF_{\theta}(x,z) + b)$$

Choose the values of the parameters α and θ such that w is maximized (maximum margin algorithm) and T, the model selection measure, is minimized (best kernel parameters). More precisely, for θ fixed,

$\alpha^{0} = \arg\max w(\alpha)$ *and choose* θ^{0} Such that

$$\theta^{0} = \arg\min_{\theta} T(\alpha^{0},\theta$$

When,θ is a one dimensional parameter, one typically try a finite number of values and picks the one which gives the lowest value of the criterion T.When both T and the SVM result are continuous with respect to h a better approach. They used an incremental optimization algorithm, one can train an SVMwith little effort when θ is changed by a small amount. However, as soon as has more than one component computing T(α, θ) for every possible value of h becomes inflexible, and one rather looks for a way to optimize θ along a trajectory in the kernel parameter space. In this work, the gradient of a model selection criterion to optimize the model parameters are used.

4. Experimental Results

In this chapter the effectiveness of the proposed DCBFS is investigated. the results of the proposed and existing system Relief-KNN, Relief –SVM, P Relief –KNN, P Relief –SVM, Map Relief –KNN, Map Relief –SVM, DCBFS-SVM, and DCBFS-GRPF. We compare all of these methods with Chess, Heart and Segment dataset. Each and every method shows accuracy when compared to all feature selection, DCBFS-GRPF shows best accuracy than other methods.

Heart Dataset
The Heart dataset describes diagnosing of cardiac Single Proton Emission Computed Tomography (SPECT) images. Each of the patients is classified into two categories: normal and abnormal. The database of 267 SPECT image sets (patients) was processed to extract features that summarize the original SPECT descriptions. As a result, 44 continuous feature patterns were created for each patient. The pattern was further processed to obtain 22 binary feature patterns.

Chess Dataset
The dataset is divided into a training dataset, representing a consecutive stretch of (for.eg) 100 months of game-by-game results among those top players, and a test dataset, representing after that 5 months of games played among those players (obviously the actual game-by-game results on the test dataset have been withdrawn). Finally find the result of the best players.

Table 4.1: Characteristics of 3 UCI Data Sets

Dataset	*Train Size*	*Test Size*	*#Feature*	*#Class*
Heart	80	187	22	2
Chess	1586	1001	34	2
Segment	170	100	13	2

4.1 Performance Evaluation Parameters

- The *accuracy* (*AC*) is the proportion of the total number of predictions that were correct. It is determined using the equation:

$$AC = \frac{a+d}{a+b+c+d}$$

- *a* is the number of correct predictions that an instance is negative,
- *b* is the number of incorrect of predictions that an instance is positive,

- *c* is the number of incorrect of predictions that an instance negative, and
- *d* is the number of correct of predictions that an instance is positive.

4.2 Clustering Based Feature Selection Performance

Discriminative clustering based feature selection; it involves the construction of the minimum spanning tree from a weighted complete graph; the partitioning of the MST (minimum spanning tree) into a forest with each tree representing a cluster; the selection of representative features from the clusters.

4.2.1 Classification Performance

For Heart, Chess, and Segment Data Set, the randomly selected data's are used for training/validation and testing data's are used for testing. Note that the data for training are separate from the data for testing in each case. Finally compare the class labels for both training data and testing data and find the accuracy of each feature. Kernel function called Gaussian Radial basis Polynomial Function (GRPF) is introduced that could improve the classification accuracy of Support Vector Machines (SVMs) for both linear and non-linear data sets.

Figure1 clearly shows the accuracy of each selected-classifier combination, as a function of the number of top-ranked features on testing from the Heart Dataset. It is clear that the discriminative clustering based feature selection with Gaussian radial based polynomial function achieve better performance than existing methods.

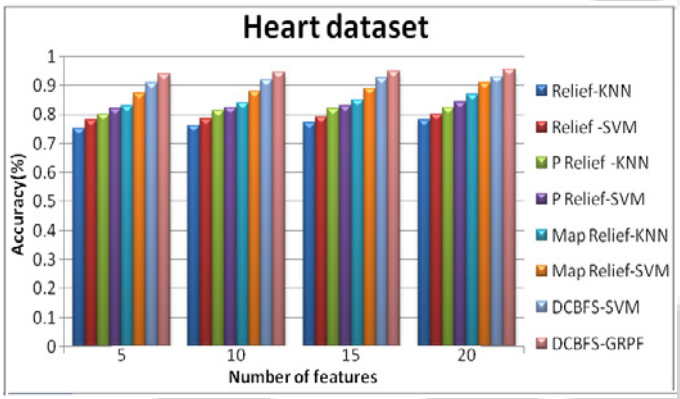

Figure 1: Comparison of Relief, P-Relief, MAP-Relief and DCBFS on Heart Dataset: testing accuracy versus the number of selected features.

Figure2 clearly shows the accuracy of each selected-classifier combination, as a function of the number of top-ranked features on testing from the Chess Dataset. It is clear that the discriminative clustering based feature selection with Gaussian radial based polynomial function achieve better performance than existing methods.

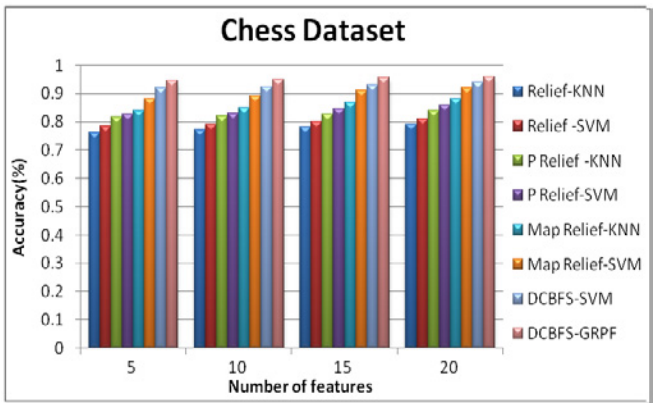

Figure 2: Comparison of Relief, P-Relief, MAP-Relief and DCBFS on Chess Dataset: testing accuracy versus the number of selected features.

Figure3 clearly shows the accuracy of each selected-classifier combination, as a function of the number of top-ranked features on testing from the Segment Dataset. It is clear that the discriminative clustering based feature selection with Gaussian radial based polynomial function achieve better performance than existing methods.

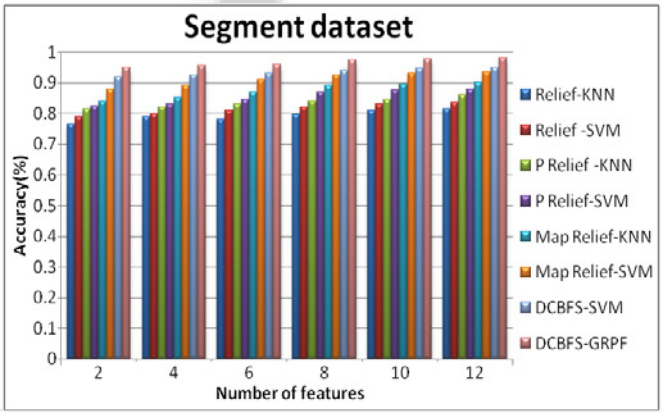

Figure 3: Comparison of Relief, P-Relief, MAP-Relief and DCBFS on Segment Dataset: testing accuracy versus the number of selected features.

Figure4 clearly shows the time comparison of Relief, P-Relief, MAP-Relief and DCBFS. The proposed DCBFS is proved the best performance of the execution time.

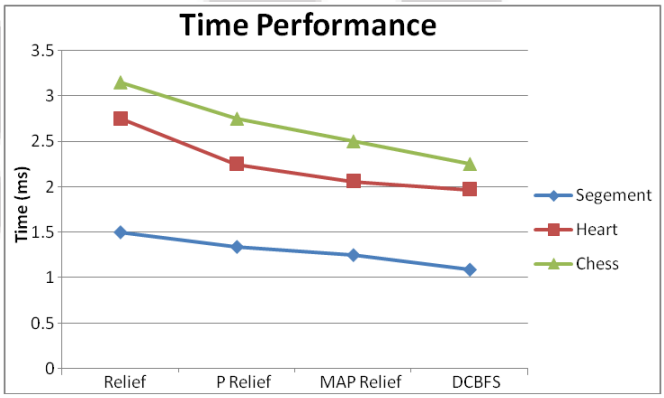

Figure 4: Time Comparison of Relief, P-Relief, MAP-Relief and DCBFS on Heart, Chess and Segment Dataset

5. Conclusion

In this research feature selection algorithm with clustering is performed to efficiently select best feature subset selection in the high dimensional data. Discriminative clustering based feature selection algorithm which possesses several compelling merits compared with its representative counterpart. After that clustering based feature subset selection was performed then apply SVM classification algorithm to high dimensional data. It shows that proposed discriminative clustering based feature subset selection with SVM best classification accuracy then the previous work. Discriminative clustering based feature selection algorithm with SVM theoretically optimal and computationally efficient. Support vector machines with the proposed new kernel function (GRPF) accomplishes better accuracy than SVM, especially in high dimension data sets .The proposed GRPF kernel has achieved the best accuracy, particularly with the data sets with many attributes.

6. Future Work

In future work DoC can be employed for learning affinity graphs or pair wise similarities and developed an algorithm for learning feature transformation. In further exploited the DoC framework and ranking aggregation.

References

[1] Veeraswamy, 2011. "A Survey of Feature Selection algorithm in Data Mining", Vol.1, Issue 2, pp-108-117.
[2] Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," J. Machine Learning Research, vol. 3, pp. 1157-1182, 2003.
[3] M.A. Hall and G. Holmes, "Benchmarking Attribute SelectionTechniques for Discrete Class Data Mining," IEEE Trans. Knowledge and Data Eng., vol. 15, no. 6, pp. 1437-1447, Nov./Dec. 2003.
[4] Y. Saeys, I. Inza, and P. Larranaga. A Review of Feature Selection Techniques in Bioinformatics. Bioinformatics, 23(19):2507–2517, 2007.
[5] K. Kira and L.A. Rendell, "A Practical Approach to Feature Selection," Proc. Ninth Int'l Workshop Machine Learning (ICML '92), pp. 249-256, 1992.
[6] Y. Sun, "Iterative Relief for Feature Weighting: Algorithms, Theories, and Applications," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 29, no. 6, pp. 1035-1051, June 2007.
[7] R. Gilad-Bachrach, A. Navot, and N. Tishby, "Margin Based Feature Selection—Theory and Algorithms," Proc. 21st Int'l Conf. Machine Learning (ICML '04), 2004.
[8] Chen, H. Liu, J. Chai, and Z. Bao, "Large Margin Feature Weighting Method via Linear Programming," IEEE Trans. Knowledge Data Eng., vol. 21, no. 10, pp. 1475-1488, Oct. 2009.
[9] K.E. Hild, D. Erdogmus, K. Torkkola, and J.C. Principe, "Feature Extraction Using Information-Theoretic Learning," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 28, no. 9, pp. 1385- 1392, Sept. 2006.
[10] N. Vasconcelos, "Feature Selection by Maximum Marginal Diversity: Optimality and Implications for Visual Recognition," Proc. IEEE CS Conf. Computer Vision and Pattern Recognition (CVPR '03), pp. 762-772, 2003.
[11] Kai Yu, X. Xu, M. Ester, and H.-P. Kriegel, "Feature Weighting and Instance Selection for Collaborative Filtering: An Information- Theoretic Approach_," Knowledge and Information Systems, vol. 5, no. 2, pp. 201-224, 2003.
[12] http://web.ist.utl.pt/ acardoso/datasets/.
[13] http://www.cs.technion.ac.il/ ronb/thesis.html.
[14] http://www.daviddlewis.com/resources/testcollections/reuters21578/
[15] http://www.dmoz.org/

Author Profile

K. Saranya received the Bachelor's degree in Computer Science from Bharathiar University in 2009.She received the Master's degree in Computer Science from Bharathiar University in 2011.

T. Deepa working as a Assistant professor in Sri Ramakrishna College of Arts and Science for women, Bharathiar University, Coimbatore, Tamilnadu. She has guided several PG and Research projects. She has presented her papers in International Conferences and has published papers in International Journals.