A NOVEL PERFORMANCE OF SUPERVISED ATTRIBUTE CLUSTERING OF DIMENSIONALITY REDUCTION IN IRRELEVANT DATA

[1]Bujji Babu Lingampalli , [2]K.S.R. Prasad, [3]E Rama Lakshmi, [4]B Nandana Kumar
[1]Assistant Professor, Department of CSE, DNR College of Engineering & Technology, Bhimavaram, AP, India.
[2]Assistant Professor, Department of CSE, DNR College of Engineering & Technology, Bhimavaram, AP, India, [3]Assistant Professor, Department of CSE, DNR College of Engineering & Technology, Bhimavaram, AP, India. [4]Assistant Professor, Department of CSE, DNR College of Engineering & Technology, Bhimavaram, AP, India.

Abstract : Feature subset selection is an effective way for dimensionality reduction, eliminating irrelevant data and redundant data, increasing accuracy. There are various feature subset selection methods in machine learning applications and they are classified into four categories: Embedded, wrapper, filter and hybrid approaches. Embedded approach is more efficient than other three approaches. Example for this approach is traditional machine learning algorithms such as decision trees and neural networks. Wrapper method gives more accuracy in learning algorithms. But here the computational complexity is large. This paper centers on a novel data mining technique we term supervised clustering. Unlike traditional clustering, supervised clustering assumes that the examples are classified. The goal of supervised clustering is to identify class-uniform clusters that have high probability densities. A novel approach called supervised attribute clustering algorithm is proposed to improve the accuracy and check the probability of the patterns. In this method, faster retrieval of relevant data is made more efficient and accurate. By using this method, users can get precise results and negligible data loss. This method displays results based on the high probability density thereby providing privacy for data and reducing the dimensionality of the data.

IndexTerms - **Embedded, Wrapper, Clustering, Hybrid, Virginica, Setosa**

I. INTRODUCTION

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis , used in many fields, including machine learning, image analysis, information retrieval bio informatics data compression and computer graphics. Generally on the unsupervised learning framework clustering is applied using particular error functions, e.g. an error function that minimizes the distances inside a cluster keeping clusters tight. The difference between Supervised and traditional clustering, is traditional clustering that is applied on classified examples with the objective of identifying clusters that have high probability density with respect to a single class.

The main advantage with supervised clustering, is we also like to keep the number of clusters small, and objects are assigned to clusters using a notion of closeness with respect to a given distance function. Fig. 1 examines the differences between traditional and supervised clustering. Let us consider that the black and white examples in the figure that represent subspecies of Iris plants named Setosa and Virginica, respectively. Here we apply traditional clustering algorithm, that identifies the four clusters depicted in Figure 1.a. If our aim is to generate summaries for the Virginica and Setosa classes of the Iris Plants, the clustering in Figure 1.a would not be very attractive because it joined Setosa and Virginica objects in cluster A whereas it com-bined the examples of Virginica class in two different clusters B and C.

The proposed supervised clustering algori- thm maximizes class purity, as well as, it splits cluster A into two clusters E and F. simultaneously the supervised clustering tries to keep the number of clusters low. Here clusters B and C would be merged into one cluster without compromising class purity where as reducing the number of clusters is the another important feature in this algorithm. Supervised clustering algorithm would finally identifies cluster G with the combination of clusters B and C as shown in Figure1.b..

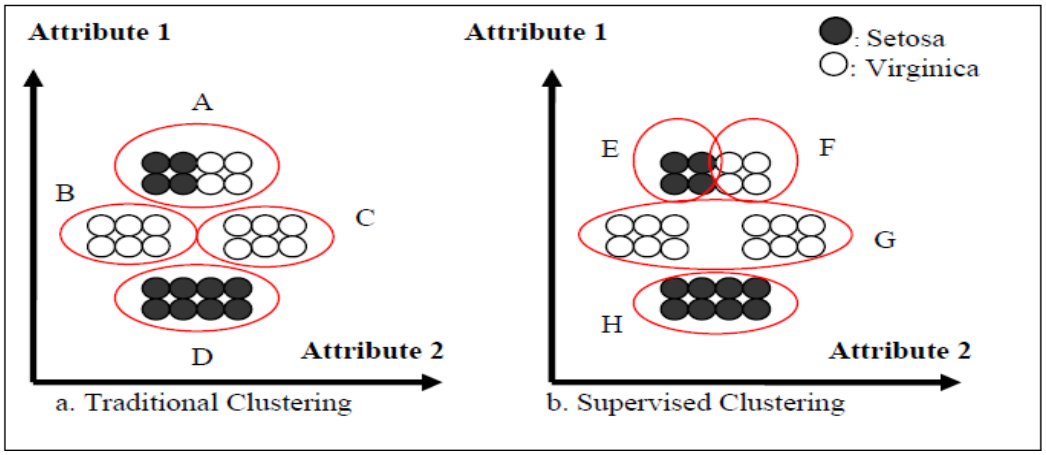

figure 1: differences between traditional clustering and supervised clustering

II. RELATED WORK

There has been some work that has some similarity with our research under the heading of semi-supervised clustering. The idea of semi-supervised clustering is to enhance a clustering algorithm by using side information in the clustering process that usually consists of a "small set" of classified examples; the objective of the clustering process, then, is to optimize classpurity (examples with different class labels should belong to different clusters) in addition to the traditional objectives of a clustering algorithm. The existing research on semi-supervised clustering can be subdivided into 2 major groups: similarity-based methods and search-based methods (for more details see [BBM03]).

Similarity-based methods create a modified distance function that incorporates the knowledge with respect to the classified examples and use a traditional clustering algorithm to cluster the data. Search-based methods, on the other hand, modify the clustering algorithm itself but do not change the distance function. [XNJ03] (and similarly [BHSW03]) take the classified training examples and transform those into constraints (points that are known to belong to different classes need to have a distance larger than a given bound) and derive a modified distance function that minimizes the distance between points in the data set that are known to be similar with respect to these constraints using classical numerical methods. The K-means clustering algorithm in conjunction with the modified distance function is then used to compute clusters. Klein [KKM02] proposes a shortest path algorithm to modify a Euclidian distance function based on prior knowledge. Demiriz [DBE99] proposes an evolutionary clustering algorithm in which solutions consist of k centroids and the objective of the search process is to obtain clusters that minimize (the sum of) cluster dispersion and cluster impurity. Cohn [CCM00] modifies the popular EM algorithm so that it is capable of incorporating similarity and dissimilarity constraints.

Finally, Basu et. al. [BBM02] centers on modifying the k-means clustering algorithm to cope with prior knowledge. However, there are two approaches that can be viewed as supervised clustering approaches. Sinkkonen et al., [SKN02], propose a very general approach called discriminative clustering that minimizes distortion within clusters. Distortion, in their context, represents the loss of mutual information between the auxiliary data (e.g., classes) and the clusters caused by representing each cluster by a prototype. The technique seeks to produce clusters that are internally as homogeneous as possible in conditional distributions p(c|x) of the auxiliary variable, i.e., belong to a single class. Similarly, Tishby et. al. introduced the information bottleneck method [TPB99]. Based on that method, they proposed an agglomerative clustering algorithm, [ST99], that minimizes information loss with respect to P(C|A) with C being a class and A being an attribute.

III. OBJECTIVES

A feature selection algorithm may be evaluated from both the efficiency and effectiveness points of view. While the efficiency concerns the time required to find a subset of features, the effectiveness is related to the quality of the subset of features. Based on these criteria, a fast clustering-based feature selection algorithm (FAST) is proposed and experimentally evaluated in this paper. The FAST algorithm works in two steps. In the first step, features are divided into clusters by using graph-theoretic clustering methods. In the second step, the most representative feature that is strongly related to target classes is selected from each cluster to form a subset of features.

Features in different clusters are relatively independent; the clustering-based strategy of FAST has a high probability of producing a subset of useful and independent features. To ensure the efficiency of FAST, we adopt the efficient minimum-spanning tree (MST) clustering method. The efficiency and effectiveness of the FAST algorithm are evaluated through an empirical study. The disadvantage with the FAST algorithm is the generality of the selected features is limited and the computational complexity is large ,and the accuracy of the learning algorithms is not guaranteed.

IV. METHODOLOGY

The proposed supervised attribute clustering algorithm is based on measuring the similarity between attributes using the new quantitative measure, whereby redundancy among the attributes is removed. A supervised clustering algorithm that maximizes class purity, the clusters are then refined incrementally based on sample categories By using this method, users can get precise results and negligible data loss. This method displays results based on the high probability density thereby providing privacy for data and reducing the dimensionality of the data

The Supervised attribute clustering algorithm starts by selecting a randomly number of examples from the given dataset as the initial set of representatives. Clusters are, created to the cluster of their closest representative by assigning examples. Algorithm Starts from the randomly generated set of representatives, and tries to improve the quality of the clustering by adding a single non-representative example to the set of representatives in addition to that by eliminating a single representative from the set of representatives. And the algorithm finally terminates if the solution quality (measured by q(X)) does not show any improvement in the fitness function. As well as, we think that the algorithm is run r (input parameter) times starting from a randomly generated initial set of representatives every time, reporting the best of the r solutions as its final outcome. The algorithm that was used for the evaluation of supervised attribute clustering is given below. It should be determined that the number of clusters k is not fixed for this algorithm and searches for "best" values of k.

ALGORITHM

REPEAT r TIMES
val:=a randomly created set of representatives(with size between c+1 and 2c)
WHILE NOT DONE DO
1.Develop new solutions S by adding a single non-representative to val and by removing a single representative from val.
2.Compute the element s in S for which q(s) is minimal,(if there are more elements than one minimal element, randomly pick one).
3.IF q(s)<q(val) THEN val=s
ELSE IF q(s)=q(val) AND mod(s)>mod(val) THEN val=s
ELSE terminate and return val as the solution for this run.
Notice the best out of r solutions found.

The Notations Used in this algorithm are O={o1, …, on} Objects in the data set, n is the Number of objects in the data set, d(oi,oj) is the distance between objects oi & oj, c is the number of classes, Ci is the cluster along with the i-th representative, X={C1, …, Ck} is the clustering solution that contains clusters of C1 to Ck, k=|x| is the number of clusters (or representatives) in a solution X, q(X) is the fitness function that determines the clustering. X.

1. The algorithm starts with a randomly generated set of representatives,
2. Firstly, the algorithm creates clustering obtained by adding a single non-representative to the current set of representatives.
3. Secondly, the algorithm creates clustering obtained by eliminating a single representative from the current set of representatives.
4. Clustering is then evaluated, and the solution whose clustering has the lowest value with respect to q(X) is selected.
5. The process continues iterating ,still there is an improvement in fitness function q(X) and therefore finds a clustering that uses small number of clusters.
6. Moreover, the algorithm did not stop, even though the class purity did not improve any further. Because fitness function q(X) does not only try to maximize the class purity, but also minimizes the number of clusters.

V. DATA SET

Here we considered input as Iris plant dataset that consists of 150 flowers, numbered 1 through 150. The subspecies of Iris plants are named as Setosa,Versicolor,and Virginica with their corresponding sepal-length, sepal-width, petal-length, petal-width respectively. The input dataset is classified into two different data groups based on assumed threshold value.

VI. RESULTS

We have applied the supervised attribute clustering algorithm to a benchmark consisting of Iris plant data set obtained from UCI Machine Learning Repository. The algorithm is evaluated based on cluster purity, value of the fitness function q(X), average dissimilarity between all objects and their representatives. value. This algorithm works starts by calculating its fitness function. The search continues using as the new set of representatives until it finds a best solution, leading to an improvement in fitness from 0.054 to 0.043. The algorithm continues iterating as long as there is an improvement in fitness function q(X).Then it computes the class purity with the set of representatives and still achieved class purity as 0.9667.Until unless the algorithm did not stop because of fitness function q(X) .It increases the class purity, and also reduces the number of clusters. Therefore it computed the redundant datasets as 16, and removes from the dataset and marks the class purity as 0.9667.

The whole process is explained with a list of figures as shown below. Among them the first diagram illustrates that how we are taking the input dataset and running.

figure 2: uploading a dataset

The second figure explains about how the algorithm is running and removing the redundant datasets.

figure 3: applying algorithm

Third figure computes the class purity of the dataset.

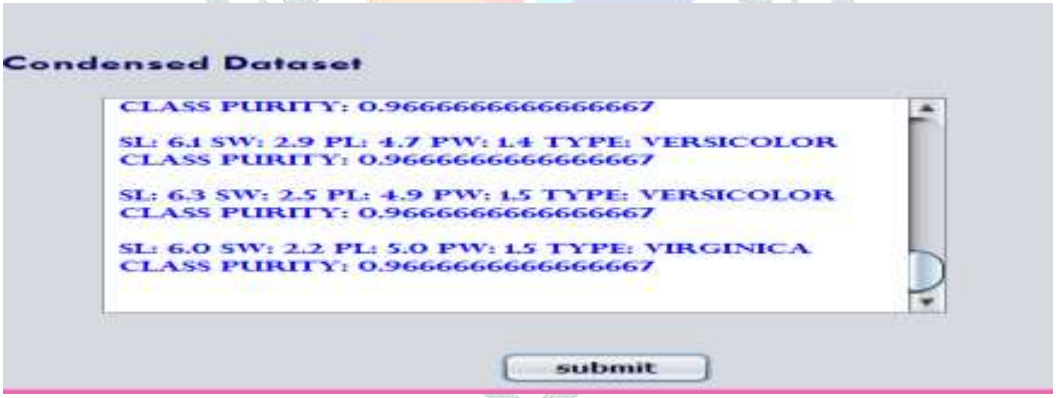

figure 4: class purity of dataset

VII. CONCLUSION

In this paper a novel data mining technique named supervised attribute clustering was introduced. Supervised attribute clustering aims to identify class-uniform clusters that have high probability density. The algorithm involves (1)removing irrelevant and redundant data.(2)computes the fitness function.(3)evaluating the class purity. Furthermore, our approach relies on distance metrics and representative-based clustering, whereas their approach is probabilistic and based on mutual information maximization.

REFERENCES

[1] Chanda P., Cho Y., Zhang A. and Ramanathan M., Mining of Attribute Interactions Using Information Theoretic Metrics, In Proceedings of IEEE international Conference on Data Mining Workshops, pp 350-355, 2009.

[2] Chikhi S. and Benhammada S., ReliefMSS: a variation on a feature ranking ReliefF algorithm. Int. J. Bus. Intell. Data Min. 4(3/4), pp 375-390, 2009.

[3] Cohen W., Fast Effective Rule Induction, In Proc. 12th international Conf. Machine Learning (ICML'95), pp 115-123, 1995.

[4] Dash M. and Liu H., Feature Selection for 4.Classification, Intelligent Data Analysis, 1(3), pp 131-156, 1997.

[**5]** Dash M., Liu H. and Motoda H., Consistency based feature Selection, In Proceedings of the Fourth Pacific Asia Conference on Knowledge Discovery and Data Mining, pp 98-109, 2000.
[**6]** Das S., Filters, wrappers and a boosting-based hybrid for feature Selection, In Proceedings of the Eighteenth International Conference on Machine Learning, pp 74-81, 2001.
[7] M.Ackerman and S. Ben-David. Clusterability: A theoretical study. Proceedings of AISTATS-09, JMLR: W&CP, 5:1–8, 2009.
[8] Ben-David S. Ackerman, M. and D. Loker. Characterization of Linkage-based Clustering. COLT 2010, 2010.
[9] D. Angluin. Queries and concept learning. Machine Learning, 2:319–342, 1998. [BB08] Maria-Florina Balcan and Avrim Blum. Clustering with interactive feedback. In ALT, 2008.
[11] Eick, C., Rouhana, A., Chen, C., Bagheriran, A., Vilalta, R. "Using Clustering to Learn Distance Functions for Supervised Similarity Assessment", submitted for publication.
[12] "Using Representative-Based Clustering for Nearest Neighbor Dataset Editing". *Æ* ICDM'04
[13] Klein,D., Kamvar,S.-D., Manning, C. "From instance-level Constraints to Space-level Constraints: Making the Most of Prior Knowledge in Data Clustering", in Proc. ICML'02, Sydney, Australia.
[14] Kaufman L. and Rousseeuw P. J. "Finding Groups in Data: an Introduction to Cluster Analysis", John Wiley & Sons, 1990.