International Journal of Computer Applications Technology and Research
Volume 3– Issue 11, 706 - 710, 2014, ISSN: 2319–8656

Feature Selection Algorithm for Supervised and Semisupervised Clustering

S.Gunasekaran
Department of Computer Science and Engineering
V.S.B.Engineering College,
Karur, India

I. Vasudevan
Department of Computer Science and Engineering
V.S.B.Engineering College,
Karur, India

Abstract−In clustering process, semi-supervised learning is a tutorial of contrivance learning methods that make usage of both labeled and unlabeled data for training - characteristically a trifling quantity of labeled data with a great quantity of unlabeled data. Semi-supervised learning cascades in the middle of unsupervised learning (without any labeled training data) and supervised learning (with completely labeled training data). Feature selection encompasses pinpointing a subsection of the most beneficial features that yields well-suited results as the inventive entire set of features. A feature selection algorithm may be appraised from both the good organization and usefulness points of view. Although the good organization concerns the time necessary to discover a subsection of features, the usefulness is related to the excellence of the subsection of features.Traditional methodologies for clustering data are based on metric resemblances, i.e., non-negative, symmetric, and satisfying the triangle unfairness measures using graph-based algorithm to replace this process in this project using more recent approaches, like Affinity Propagation (AP) algorithm can take as input also general non metric similarities.

Keywords:Data mining,Feature selection, Feature clustering, Semi-supervised, Affinity propagation

1. INTRODUCTION

Clustering algorithms can be categorized based on their cluster model. The most appropriate clustering algorithm for a particular problem often needs to be chosen experimentally. It should be designed for one kind of models has no chance on a dataset that contains a radically different kind of models.For example, k-means cannot find non-convex clusters. Difference between classification and clusteringare two common data mining techniques for finding hidden patterns in data. While theclassification and clustering is often mentioned in the equal sniff, and dissimilar analytical approaches.

There isdiversity of algorithms rummage-sale for clustering, but all the share belongings of

Iteratively assigning records to a cluster, manipulative a quantity and re-assigning records to clusters until the designedprocedures don't modification much demonstrating that the process has converged to firmsections. Records within a cluster are more comparable to every one other, and added different from records that are in other clusters. Contingent on the precise implementation, there are a diversity of procedures of resemblance that are rummage-sale tooverallaim is for the attitude to converge to collections of correlated records.Classification is a dissimilarmethod than clustering. Classification is correlated to clustering in that it also segments customer records into distinctive segments called classes. But dissimilar clustering, a classification inquiry requires that the end-user/analyst know ahead of time how classes are demarcated.

For instance, classes can be demarcated to represent the probability that a customer nonpayment on a loan (Yes/No). It is essential that every record in the dataset rummage-sale to physique the classifier before now have a value for the trait rummage-sale to describe classes. Because every record has a value for the trait rummage-sale to describe the classes, and because the end-user resolves on the trait to use, classification is much less investigative than clustering. The impartial of a classifier is not to search the data to ascertain interesting segments, but relatively to select how new records should be classifiedi.e. is this new customer likely to default on the loan?

With the aim of selecting a subsection of good features with high opinion to the impartialperceptions, feature subsection selection is a real way for reducing dimensionality, rejectingunrelated data, inflammation learning accurateness, and purifying result unambiguous.Feature subsection selection can be observed as the progression of ascertaining and confiscating as variousunrelated and redundant features as possible. This is because 1) unrelated features do not subsidize to the extrapolation exactitude and 2) redundant features do not redound to receiving anenhancedanalyst for that they delivergenerally information which is previouslycontemporary in other feature(s). Unrelatedfeatures, beside with redundant features, strictly affect the exactness of the learning technologies.

Thus, feature subsection selection should be able to identify and remove as much of the unrelatedand redundant information as possible. Itdevelops a novel algorithm which can efficiently and effectively deal with both unrelatedand redundant features, and obtain a good feature subsection. We achieve this through a new feature selection framework which composed of the two connected components of unrelatedfeature removal and redundant feature removal. The previousacquires features relevant to the target concept by eliminating unrelatedones, and the latter removes redundant features from relevant ones via choosing denotative from different feature clusters, and thus produces the final subsection.

A fast clustering-based feature selection algorithm (FAST) works in two steps. In the first step, by using graph-theoretic clustering methods the features are separated into clusters. In the second step, the most typical feature that is powerfullyassociated to target classes is designated from every cluster to form a subsection of features. Features in different clusters are comparativelyindependent; the clustering-based approach of FAST has a high probability of producing a subsection of useful and sovereign features. To make sure the effectiveness of FAST, assume the well-organized minimum-spanning tree (MST) clustering method.

The unrelatedfeature removal is straightforward once the right relevance measure is demarcated or selected, while the redundant feature elimination is a bit of refined. Inthe FAST algorithm, it encompasses 1) the structure of the minimum spanning tree from a weighted complete graph; 2) the partitioning of the MST into a forest with every tree denoting a cluster; and 3) the selection of denotative features from the clusters.Feature selection encompassesdetecting a subsection of the most useful features that produces compatible results as the original entire set of features.

2. RELATED WORK

The proposed method [2] providesthe number of features in numerous applications where data hashundreds or thousands of features. Existing feature selection approachespredominantly focus on verdict relevantfeatures. In this feature selection display that feature relevance alone is inadequate for well-organized featureselection of high-dimensional data. We define feature redundancy and propose to perform explicitredundancy analysis in feature selection. A new framework is introduced that decouples relevanceanalysis and redundancy analysis. We develop a correlation-based method for relevance and redundancy analysis, and conduct an empirical study of its efficiency and effectiveness comparing withrepresentative methods.

The novel algorithm for discovery non-redundant discarded feature subsections based on the PRBF[5]has only one consideration, numericalmeaningor the likelihood that the assumption that disseminations of two features arecomparableis true. In the first step directories have been rummage-sale for ranking, and in thesecond step terminated features are detached in an unsupervised way, because during decrease of terminated features data about the modules is not used.

The primary tests are promising: on the reproduction data perfect ranking has been re-formedand terminated features rejected, while on the real data, with relatively modest number offeatures selected outcomes are regularly the superlative, or close to the superlative, associating withfour state-of-the-art feature selection algorithms. The novel algorithm appears to workespecially well with the directSVM classifier. Computational anxieties of PRBF algorithmare related to other correlation-based filters, and lower than Relief.

The searching for interacting features in subsectionselection [9] developing and acclimatizingabilities of robust intellect are superlativeestablished in its aptitudeto learn. Mechanism learning facilitates computer systems to learn, and recoverpresentation. Featureselection facilitatesmechanism learning by targeting to eliminate irrelevant features.Feature interaction presents a dare to feature subsection selection for cataloging.This is because a feature by itself might have little relationship with the objective concept, but whenit is combined with some other features, it can be strongly interrelated with the objective concept.

Thus, the inadvertentelimination of these features may effect in poor catalogingpresentation. It is computationally inflexible to switch feature exchanges in general. Nevertheless, the attendanceof feature interaction in anextensive range of real-world requests demands applied solutions thatcan decrease high-dimensional data although perpetuating feature exchanges. In this paper, it ups the contest to design a special data structure for feature quality evaluation, and to employ an information-theoretic feature ranking mechanism to efficiently handle feature interaction in subset selection.

We conduct experiments to evaluate our approach by comparing with some representative methods, perform a lesion study to examine the critical components of the proposed algorithm to gain insights, and investigate related issues such as data structure, ranking, time complexity, and scalability in search of interacting features.

The success of many feature selection algorithms allows us to tackle challenging real-world problems. Many applications inherently demand the selection of interacting features.

An Evaluation on feature selection for text clustering is first demonstrated that feature selection can improve the text clustering efficiency and performance in ideal case, in which features are selected based on class information. But in real case the class information is unknown, so only unsupervised feature selection can be exploited. In many cases, unsupervised feature selection are much worse than supervised feature selection, not only less terms they can remove, but also much worse clustering performance they yield.

3. PROPOSED SYSTEM

Traditional approaches for clustering data are based on metric resemblances, i.e., nonnegative, symmetric and filling the triangle disparity measures. More recent approaches, like Affinity Propagation (AP) algorithm can take as input also general non metric similarities. AP can use as input metric selected segments of images’ pairs. Accordingly, AP has been rummage-sale to solve a wide range of clustering problems, such as image processing tasks gene detection tasks, and individual preferences predictions.

Affinity Propagation is derived as an application of the max-sum algorithm in issue graph; it is used to explorations for the smallest amount of dynamism function on the basis of message passing between data points. In this system implementsthe semi supervised learning has taken a great deal of considerations. It is a mechanism learning paradigm in which the model is constructed using both labeled and unlabeled data for training set.

It retrieve the data from training data or labeled data and extract the feature of the data and compare with labeled data and unlabeled data .In clustering process, semi-supervised learning is a class of machine learning techniques that make use of both labeled and unlabeled data for training - typically a small amount of labeled data with a large amount of unlabeled data.

Semi-supervised learning cascades among unsupervised learning (without any labeled training data) and supervised learning. Various mechanism-learning investigators have found that unlabeled data, when rummage-sale in conjunction with a small amount of categorized data, can yieldsubstantialdevelopment in learning accuracy.

3.1 Irrelevant Based Feature Selection

A feature selection algorithm may be appraised from together the proficiency and usefulness point of view. Although the effectiveness concerns the time requisite to find a subsectionof features, the efficiency is associated to the excellence of the subsection of features.

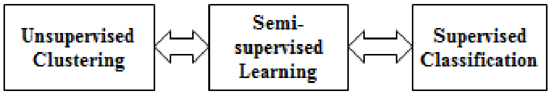

Fig 1: Semi-Supervised Learning

Many feature subsection selection algorithms, some can successfully remove irrelevant features but fail to handle redundant features yet some of the others can eliminate the irrelevant while taking care of the redundant features. In this system the FAST algorithm cascades into the subsequentgroup. Theprevious obtains features pertinent to the target concept by eliminating unrelated ones, and then removes redundant features from pertinent ones via choosing denotative from different feature clusters.

3.2 Redundant Based Feature Selection

The hybrid methods are combination of filter and wrapper methods by using a filter method to reduce search space that will be considered by the succeeding wrapper. It focuses on coalescing filter and wrapper approaches to achieve the best possible performance with a particular learning algorithm with similar time complexity of the filter methods. Redundant features do not redound to getting a better predictor for that they provide mostly information which is already present in other feature(s).

3.3 Graph Based Cluster

An algorithm to systematically add instance-level constraints to the graph based clustering algorithm. Unlike other algorithms which use a given static modeling parameters to find clusters, Graph based cluster algorithm finds clusters by dynamic modeling. Graph based cluster algorithm uses both Closeness and interconnectivity while identifying the most similar pair of clusters to be merged.

3.4 Affinity Propagation Algorithm

The affinity propagation (AP) is a clustering algorithm established on the notion of "message passing" among data points. For example of clustering algorithm is k-means. It does not need the quantity of clusters to be determined or estimated before running the algorithm.

Let x_1 and x be a set of data points, with no expectations ready around their internal structure, and the function that measures the resemblance among any two points, that is $s(x_i, x) > s(x_i, x)$ if x is further related to x_i than x.

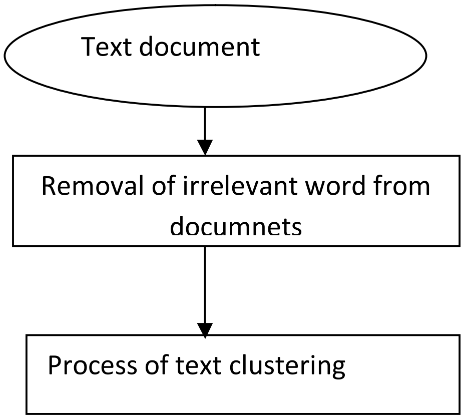
Text document
Removal of irrelevant word from documnets
Process of text clustering

Fig 2: Process of clustering

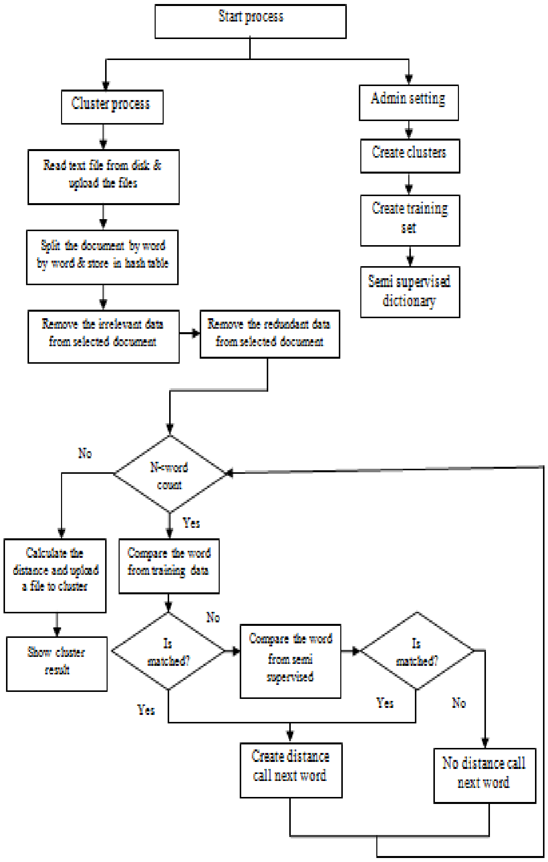
Start process
Cluster process
Admin setting
Read text file from disk & upload the files
Create clusters
Create training set
Split the document by word by word & store in hash table
Semi supervised dictionary
Remove the irrelevant data from selected document
Remove the redundant data from selected document
No
N-word count
Yes
Calculate the distance and upload a file to cluster
Compare the word from training data
No
Show cluster result
Is matched?
Compare the word from semi supervised
Is matched?
Yes
Yes
No
Create distance call next word
No distance call next word

Fig 3: system flow diagram for proposed system

The algorithm ensues by flashing two message passing steps, it modernize by using the subsequent twoconditions:

- The "responsibility" conditions*R* has values *r*(j, n) that measure how well-matched*x* is to aid as the exemplar for *x*, comparative to other candidate exemplars for *x*.
- The "availability" conditions*A* contains values *a*(j, n) characterizes how "applicable" it would be for *x* to pick *x* as its exemplar, taking into interpretation other points' favorite for *x* as an exemplar.

Together conditions are reset to all zeroes, and can be regarded as probability counters. The following updates are iteratively used to perform the algorithm:

First, responsibility updates are sent around:

$$r(j,n) \leftarrow s(j,n) - \max_{n' \neq n}\{a(j,n') + s(j,n')\}$$

Then, availability is updated per

$$a(j,n) \leftarrow \min\begin{pmatrix} 0, r(n,n) + \\ \sum_{j' \epsilon \{j,n\}} \max\left(0, r(j',n)\right) \end{pmatrix}$$

$$for^{\ j \neq n}\ and$$

$$a(n,n) \leftarrow \sum_{j' \neq n} max\left(0, r(j',n)\right)$$

4. EXPERIMENTAL RESULTS

The performance of the proposed algorithm is compared with the two well-known feature selection algorithms FCBF and CFS of text data from the aspects of the proportion of selected features and runtime analysis.

TABLE 1Runtime (in ms) of the Feature Selection Algorithms

Data set	FAST (Affinity Propagation)	FAST (Graph Based)	FCBF	CFS
Chess	90.1	94.02	94.02	90.43
Elephant	95.35	98.09	99.94	99.97
Wap.wc	69.01	71.25	75.74	77.8
Colon	87.4	90.45	90.76	89.14
GCM	55.69	58.73	59.16	60.92
AR10P	74.05	77.69	75.54	79.54
B-cell1	79.21	81.01	82.94	87.33

The affinity propagation algorithm is used to reduce the runtime compare with the graph based algorithm of FAST. It reduces the error and simplicity of performance. The semi-supervised learning is a tutorial of contrivance learning methods that make usage of both labeled and unlabeled data for training - characteristically a trifling quantity of labeled data with a great quantity of unlabeled data.

It is used to improve the efficiency of feature selection of FAST algorithm. Affinity propagation algorithm is used to achieve good performance of processing time. It provides better results with less amount of time compare with graph based algorithm.

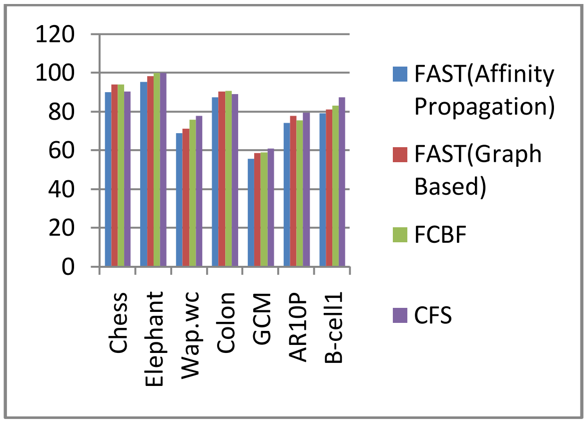

Fig 3: Runtime (in ms) of the Feature Selection Algorithms

5. CONCLUSION

In this paper, the semi supervised learning retrieve the data from training data or labeled data and extracts the feature of the data and compare with labeled data and unlabeled data. Feature selection encompasses pinpointing a subsection of the most beneficial features that yields well-suited results as the inventive entire set of features. A feature selection algorithm may be appraised from both the good organization and usefulness points of view. Then we use Affinity propagation algorithm for low error, high speed, flexible, and remarkably simple clustering algorithm that may be rummage-sale in forming teams of participants for business simulations and experiential exercises, and in organizing participants' preferences for the parameters of simulations.

5. REFERENCES

[1] Qinbao Song, Jingjie Ni, and Guangtao Wang, "A Fast Clustering-Based Feature Subset Selection Algorithm for High-Dimensional Data"*IEEE Transactions on knowledge and data engineering*vol. 25, no. 1, January 2013.

[2]L. Yu and H. Liu, "Efficient Feature Selection via Analysis ofRelevance and Redundancy," *J. Machine Learning Research*, vol. 10,no. 5, pp. 1205-1224, 2004.

[3]C.Sha, X.Qiu, and A.Zhou, "Feature Selection Based on a NewDependency Measure," *Proc. Fifth Int'l Conf. Fuzzy Systems and KnowledgeDiscovery*, vol.1,2008..

[4] I.S.Dhillon, S.Mallela, and R.Kumar, "A Divisive Information Theoretic Feature Clustering Algorithm for Text Classification,"*Machine Learning Research*, vol. 3, 2003.

[5] J.Biesiada and W.Duch, "Features selection for High-Dimensional data a Pearson Redundancy Based Filter," *Advances in Soft Computing*, vol. 45, 2008.

[6] P.Chanda, Y.Cho, A.Zhang, and M.Ramanathan, "Mining of Trait Interactions Using Information Theoretic Metrics," *Proc. IEEE Int'l Conf. Data Mining Workshops*, 2009.

[7] S.Chikhi and S.Benhammada, "ReliefMSS: A Variation on a Feature Ranking Relief Algorithm," *Int'l J. Business Intelligence and Data Mining,* vol. 4, nos. 3/4, 2009.

[8] S.Garcia and F.Herrera, "An Extension on Statistical Comparisonsof Classifiers over Multiple Data Sets for All PairwiseComparisons," *J. Machine Learning Res.*, vol. 9, 2008.

[9] Z.Zhao and H.Liu, "Searching for Interacting Features in Subset Selection," *J. Intelligent Data Analysis*, vol. 13, no. 2, 2009.

[10]Z. Zhao and H. Liu, "Searching for Interacting Features," Proc.20th Int'l Joint Conf. *Artificial Intelligence*, 2007.