World Journal of Engineering Research and Technology
WJERT
www.wjert.org
SJIF Impact Factor: 3.419

SET COVER-RRK FEATURE SELECTION TECHNIQUE FOR ENHANCING THE ACCURACY IN SOCIAL NETWORK DATA

R.Rajkumar[1*] and Dr. Anbuselvi[2]

[1,2]Bishop Heber College, Trichy.

Article Received on 29/10/2016 **Article Revised on 18/11/2016** **Article Accepted on 08/12/2016**

***Corresponding Author**
R.Rajkumar
Bishop Heber College, Trichy.

ABSTRACT

Online Social Network like Face book, Twitter, LinkedIn etc., have become the popular interaction, recreation and socialization facility on the internet. Users choose greater engaging sites, every time they will notice familiar faces like friends, relatives or colleagues. A "feature" or "attribute" or "variable" refers to an issue of data. Usually earlier than gathering data, features are detailed or chosen. Features may be discrete, continuous, or nominal. Generally, features are characterized as: 1. Relevant: There are features which have a power on the output and their position cannot be assumed by using the relaxation. 2. Irrelevant: Irrelevant features are defined as those features not having any influence on the output, and whose values are generated at random for every instance. Feature subset selection in Online Social Network can be analyzed as the exercise of identifying and removing of as lot of irrelevant and unnecessary features as achievable. This is for the cause that, irrelevant features do no longer make a contribution to the predictive accuracy. First shifting out irrelevant features from the Online Social Network data set[5], for irrelevant features are removed by using the features having the value above the predefined threshold. The reason of this research paper is twofold; Identifying and removing the irrelevant features in Online Social Network with latest solutions for Consistency Measure in an Online Social Network.

KEYWORDS: Online Social Network, irrelevant features, relevant features, Consistency Measure.

1. INTRODUCTION

Online Social Network represents an emerging area which also brings many challenges and research opportunities besides numerous socializing facilities. This research paper focuses on the Consistency Measure for the irrelevant feature. Feature sub set selection is a valuable way for dropping dimensionality, removing irrelevant data, rising learning correctness, and improving result clarity.[1] Data mining implements various algorithms on such data tries to dig useful information by looking at a small fraction of a large amount of data.

2. FEATURE SELECTION PROCESS

Feature Selection is to find a subset of features according to the given evaluation criterion. Each feature subset is a feature combination. By evaluating each selected subset, we can reduce the number of possible feature combinations by eliminating the irrelevant features and thus simplifying the classifier.[2]

2.1 Traditional feature selection process

Traditionally, feature selection consists of four components: a subset generation system, an evaluation function or criterion and a validation procedure. Subset era process typically makes use of certain looking method to produce candidate feature subset. Each selected subset is evaluated with the aid of a criterion for its advantage and is as compared with the previous best result. If the new decided on subset has better advantage than the preceding first-rate end result the previous subset is update through the new subset. The process of subset generation and evaluation is repeated until a stopping criterion is met.

3. FEATURE SELECTION MODELS

Feature Selection algorithms designed with different evaluation standards normally fall into three categories: Filter model, Wrapper model[3] and Hybrid model. The Filter model evaluates feature subsets based totally on preferred characteristics of statistics without regarding any algorithm. The Wrapper model calls for evaluation criterion with a predetermined learning algorithm. The Hybrid model is a combination exploiting benefits from both Filter and Wrapper model.

4. EVALUATION CRITERIONS

Evaluation Criterions are focused on choosing relevant features and eliminating irrelevancy and redundancy. The definitions of feature relevance are categorized into three classes: strong relevant, weakly relevant and Irrelevant.[4] If a feature is strongly relevant, it shows that the

feature is to decide an optimal subset. Removing it will affect the class distribution. Weak relevant features are not always needed to attain optimal subset. In this paper we recognition on a well-known evaluation criterion known as consistency measure.

4.1 Inconsistency Rate

To be consistent and concise, we denote $CCON$ as the consistent count, INC as the inconsistent count, $CCONR$ as the consistency rate, $INCR$ as the inconsistency rate.

The aim of feature selection is to find as minimal as feasible some of feature subsets which could always discriminate classifier as though the usage of the full set of features. The consistency measure as an evaluation criterion is used to decide which feature ought to be eliminated and thus decreasing the size of the feature set. The consistency rate is described through the inconsistency rate where two instances are taken into consideration inconsistent in the event that they have the same feature values however different class labels. To compute the inconsistency rate, first we must compute inconsistency count.

The inconsistency rate is the summation of all the inconsistent counts overall patterns divided by total number of instances in the dataset. The inconsistency rate $INCR$ can be expressed as

$$INCR = \sum_{i=1}^{h} INCi \div M$$

The inconsistency rate is applied into the search algorithms. A threshold R is usually defined at the start. For every feature subset F decided on by the search algorithm, the inconsistency rate $INCR$ is calculated. If the end result meets the condition $INCR \leq R$, then the subset F is taken into consideration to be consistent. The original threshold R is updated by better results. Otherwise, the subset is removed in conjunction with its features.

4.2 Consistency Rate

The Consistency rate is similar to inconsistency rate except that consistent count $CCON$ is computed.

$$CCONR = \sum_{i=1}^{h} CCONi \div M$$

5. HEURISTIC SEARCH

There are two essential goals in computer algorithms: finding a way to use a shorter running time and to produce a most advantages solution. A heuristic algorithm is used while there is no recognized way to find an optimal solution in which case the goal is to develop a simple process with a provable better running time and a stepped forward solution. Due to fact that exhaustive search algorithms take a significant amount of unnecessary time and computationally costly, the heuristic algorithm is a good alternative to fast conduct and return a first rate end result.

There are many heuristic search techniques in exercise consisting of $SetCoverRRK$. $SetCoverRRK$ is to be the most time efficient, close to optimal. The unique idea for $SetCoverRRK$ is that two instances with different class labels are said to be covered whilst there exists at least one feature which takes different values for the two instances.

Algorithm

Input: Data D, full feature set FS

Output: Consistent Feature Subset

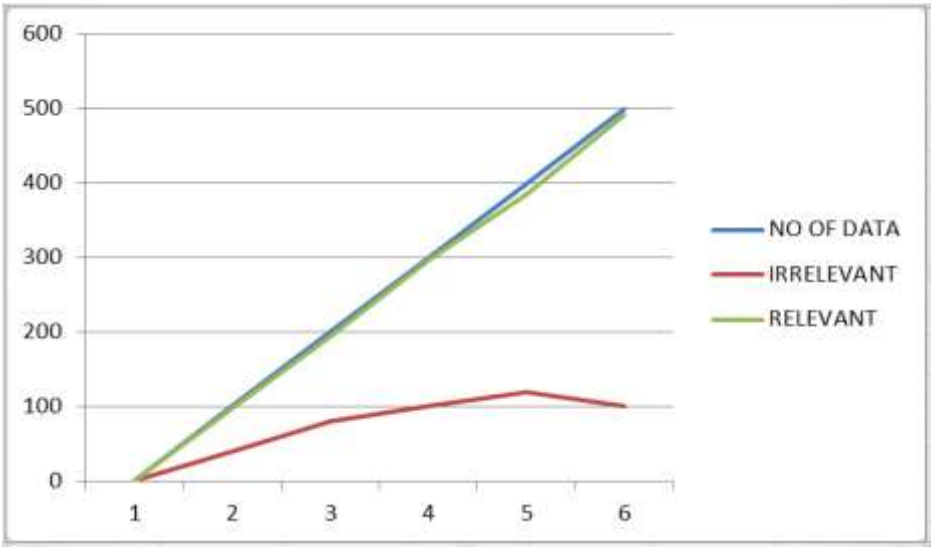
1. $BestConRate = ConCal(FS, D)$
2. $SelectedFeatureSet = [$
3. $FS' = \emptyset$
4. $LargestConRate = -\infty$
5. $\forall\ feature\ f \in FS$
   $begin$
6. $FS'' = FS' appends\ f$
7. $TempConRate = ConCal(FS'', D)$
8. $If\ TempConRate = BestConRate$
9. $Return\ FS''$
10. $Elseif\ Temp\ ConRate > LargestConRate$
11. $LargestConRate = TempConRate$
12. $SelectedFeatureSet = f$
13. $FS' = FS' append\ f$
14. $FS = FS - f$
15. end

6. DATA SET

Manage User Details

Face Book Twitter LinkedIn

	Userid	name	gender	dob	address	mobile	email	pwd	id	join_date	snw
DELETE	1	sharmi	Female	14.Jan.1993	aarathangi	8344526894	sharmi@gmail.com	sharmi	BPKfUZ	27:03:2016	FaceBook
DELETE	101	rams	Female	2.Nov.1992	madurai	9655446994	rams@gmail.com	rams	GBJL7m	27:03:2016	LinkedIn
DELETE	201	Anjali	Female	4.Mar.1982	London	7654390823	anjali@gmail.com	anjali	fiZqN9	31:03:2016	Twitter
DELETE	301	simbu	Male	4.Jan.1992	puthupatti	9977886655	simbu@gmail.com	simbu	3niVuAS	15:04:2016	FaceBook
DELETE	401	simbu	Male	4.Jan.1992	puthupatti	9977886655	simbu1@gmail.com	simbu1	qUf7ty	15:04:2016	Twitter
DELETE	501	simbu	Male	4.Jan.1992	puthupatti	9977886655	simbu2@gmail.com	simbu2	nFMdyG	15:04:2016	LinkedIn
DELETE	601	keerthi	Female	2.Jun.1992	puthukotti	9876543210	keerthi@gmail.com	keerthi	nFKRtN	27:03:2016	LinkedIn
DELETE	701	nivetha	Female	25.Aug.1993	thanjavur	8973946796	nivetha@gmail.com	nivetha	KWj85c	28:03:2016	FaceBook
DELETE	801	mounika	Female	7.Jan.1992	madathahalli	9750060023	mounika1@gmail.com	mounika1	Az4kJH	16:04:2016	Twitter
DELETE	901	manju	Female	1.May.1993	manarkudi	9988776645	manju@gmail.com	manju	gQMjVx	13:04:2016	Twitter
DELETE	1001	sharmi	Female	15.Jan.1992	madathahalli	7750724707	sharmi1@gmail.com	sharmi1	k19dZK	14:04:2016	Twitter
DELETE	1101	sharmi	Female	16.Jan.1991	salem	8887722334	sharmi2@gmail.com	sharmi2	1JAPXn	14:04:2016	LinkedIn



SETCOVER RRK Algorithm Result

The *setcoverRRK* algorithm obtains features relevant to the target concept by eliminating irrelevant features.

CONCLUSION

We present that more concept on feature subset selection, using *setcoverRRK* algorithm for calculate measuring for consistency on social network dataset. The process of *setcoverRRK*

algorithm to identifying and remove the irrelevant features and measuring for consistent rate. There are many heuristic search techniques in practice such as *setcoverRRK* is most time efficient, close to optimal. *setcoverRRK* Algorithm is used when there is no known way to find an optimal solution in which case the goal is to develop a simple process with provable better running time and an improved solution. Also *setcoverRRK* algorithm removing the irrelevant features and targeted data result can be achieved.

REFERENCES

1. Ghazi Fuad Khamis and Ramadas Naik.T., Identifying and removing irrelevant and redundant features in high dimension data using feature subset., Volume no: 2- 2015.
2. Sutha.k and Dr.Jebamalar Tamilselvi.J., A review of feature selection algorithms for data mining techniques., Volume no: 7- june-2015.
3. Randall wald, Taghi M.K hoshgoftaar, Amri Napolitano., Stability of Filter – and Wrapper - Based Feature subset selection. IEEE 25th International Conference., 2013.
4. Vipin kumar and Sonajharia minz., Feature Selection: A literature Review., June 2014.
5. Antonela tommasel., Integrating Social Network Structure in to Online Feature Selection IJCAI- 2016.