**REPORT**



**과목 : 파이썬활용고급데이터분석**

**학과 : 데이터사이언스융합학과**

**학번 : 2018711171**

**이름 : 김대겸**

**1. association rules**

. 논문명 : Software defect prediction based on correlation weighted class association rule mining

. 저자 : Yuanxun Shao, Bin Liu, Shihai Wang ∗ , Guoqi Li

**1-1. summary**

|  |
| --- |
| Software defect prediction based on supervised learning plays a crucial role in guiding software testing for resource allocation. In particular, it is worth noticing that using associative classification with high accuracy and comprehensibility can predict defects. But owing to the imbalance data distribution inherent, it is easy to generate a large number of non-defective class association rules, but the defective class association rules are easily ignored. Furthermore, classical associative classification algorithms mainly measure the interestingness of rules by the occurrence frequency, such as support and confidence, without considering the importance of features, resulting in combinations of the insignificant frequent itemset. This promotes the generation of weighted associative classification. However, the feature weighting based on domain knowledge is subjective and unsuitable for a high dimensional dataset.  Software defect prediction (SDP) has been extensively used various software attributes (also metrics or features, such as cyclomatic complexity (vg) , Chidamber and Kemerer (C&K) and process maturity ) to predict software defects before testing , which helps to reduce costs, allocate testing resources reasonably and improve software quality. Currently SDP work mainly focuses on defect proneness , defect count, defect type and defect rank. This paper explores the relationship between metrics as independent variables and binary dependent variables (defect prone (D) or non-defect prone (ND)) in historical data to predict the defect proneness of new software entities.  Although various statistical and machine learning algorithms have been investigated for SDP, such as Naive Bayes (NB), logistic regression (LR) , non-linear classifiers (e.g. support vector machines (SVM), neural network (NN)), ensemble learning (e.g. random forest (RF), bagging), tree/rule-based classifiers (e.g. OneR, RIPPER, decision tree (C4.5), decision table (DT), partial decision trees (PART)) and artificial immune system (AIS), there are still many challenges: (1) Nonlinear classifiers and ensemble learning are often called "black-box" models that sacrifice complexity to find high predictive performance, but bring poor comprehensibility which affects the performance reliability. (2) Traditional rule/tree-based classifiers (C4.5, OneR, PART) which are often regarded as "white-box" model have good understandability, but their generalization performance underperforms because they only retain partial rules for SDP.  Association rule mining can discover hidden association patterns in a dataset , and support (supp) and confidence (conf ) are two commonly used measures. It is roughly divided into two kinds of association rule mining techniques, association rule analysis (ARA) and class association rule mining (CAR). The difference between ARA and CAR lies in the constraint type of rule consequent. ARA is utilized to discover the rules consisting of features, such as basket analysis and web recommendation, whose right-hand-side only be a set of items rather than a class label. And CAR is utilized to discover classification rules consisting of features and a class label, such as classification based on associations (CBA), whose right-hand-side only be a class label rather than a set of items. And it is the integration of classification rule mining and association rule mining. Moreover, the former is unsupervised learning while the latter is supervised learning. This paper focuses on supervised CAR algorithm with an implication of the form A ⇒ C, where A is a feature subset, C is a class label. It can be interpreted as: if a sample satisfies the feature subset A, then it can be classified as C. A CAR classifier can yield a complete set of production rules with high accuracy and comprehensibility, which has attracted much attention in many domains, especially SDP.  Hence, we present a novel software defect prediction model based on correlation weighted class association rule mining (CWCAR). It leverages a multi-weighted supports-based framework rather than the traditional support-confidence approach to handle class imbalance and utilizes the correlation-based heuristic approach to assign feature weight. Besides, we also optimize the ranking, pruning and prediction stages based on weighted support. Results show that CWCAR is significantly superior to state-of-the-art classifiers in terms of Balance, MCC, and Gmean |

**1-2. reivew**

|  |
| --- |
| I think It is highly appreciated that the new software defect prediction model was presented by combining correction weights, not just basic assocation rules.  Their proposed model CWCAR is more effective than other baselines (e.g. CBA, RAAR, Naive Bayes, random forest, decision table, PART) in software defect prediction.  From the model architecture perspective, CWCAR mainly solves class imbalance and attribute weighting, and optimizes rule training. In most cases, software defect dataset is unbalanced,where D samples are less than that of ND samples. Class imbalance is at the root of poor software prediction performance. It makes associative classification find a large number of non-defective rules, and defective rules with lower support or confidence are pruned. Although associative-based classification algorithms have high accuracy, they are less able to find defects. Moreover, they use support and confidence measures to evaluate CARs without considering application areas.  In this study, we adopt multiple supports instead of support and confidence to solve the unbalanced problem. It contributes to increase the number of defective CARs by specifying different supports in accordance with the categories. This method does not employ any technique built on data-level or algorithm level, which may cause information loss or change data distribution. Different code metrics have different meanings and different importance. We incorporate attribute weighting into the association classification rule mining to resolve the indiscriminate combination and improve the quality of the rules.  Feature weights are automatically assigned using Pearson’s correlation rather than domain expert experience. Item weight, item set weight and weighted support are all calculated based on feature weight. The difference between item set weight and the previous research is that it is based on the product function of item weights instead of the average or sum function.  Besides, rule ranking, rule pruning and rule prediction are optimized based on weighted support, which are different from previous associative classification. In terms of rule ranking, the higher the weighted support, the shorter the antecedent and the earlier the lexicographic order, the higher the priority. In terms of rule pruning, redundant rule pruning based on weighted support and conflicting rule pruning are proposed. When the antecedents of any two rules are the same and the consequents are different, both rules are eliminated to avoid learning towards a certain category (D or ND).  In terms of rule prediction, the sum of multiple weighted support is utilized to classify new samples, which overcome the deficiencies of using a single rule to favor major the non-defective class.  To sum up, the proposed method CWCAR not only increases the number of defective class association rules generated, but also improves the quality of class association rules.  Through this paper, I studied more deeply about assocation rules, and also studied newly combined models as well as assocation rules, so I thought again about efficiency, scalability and new modeling.  In addition, I have confirmed and studied various techniques about how to find the law of association between data, and I think that the model proposed in this paper can cover the effects of sensitive lightness and that the algorithm's contrasting performance is very high, so I am thinking about applying it to my thesis writing and research. |

**2. social network analysis**

. 논문명 : 사회연결망 분석을 이용한 국방 강소벤처 Tech-Fi Net 기술동향 분석

. 저자 : Yuanxun Shao, Bin Liu, Shihai Wang ∗ , Guoqi Li

**2-1. summary**

|  |
| --- |
| Currently, the domestic defense industry, which is centered on domestic demand, is feared to face limitations such as slowing export growth and continue its management crisis in the defense industry. Therefore, the paradigm shift from the closed and domestic defense industries to open and export-driven ones is emphasized as a new growth engine to overcome this problem, and the government is strengthening the system to create quality jobs by fostering small and medium-sized companies and venture companies to transform them into "export-oriented defense industries." Since the growth of the sustainable defense industry begins with technological innovation of small and medium-sized defense companies and thereby securing competitiveness in the defense industry, it is necessary to lay the foundation for the active application and utilization of superior technologies of domestic small and medium-sized venture companies in the defense sector.  The purpose of this research was to analyze technological trends of Military Small Giant Venture Tech-Fi Net from Defense Technology Information Service via social network analysis. Military Small Giant Venture Tech-Fi Net, which was constituted for their fine technology for application to the military field, registered 847 technologies of 388 companies. In this research, They analyzed 847 technologies for the relations between Military System and Military Technology Category via centrality measurement, one of the social network analysis methods. The results indicate that the major technologies of domestic military small giant venture companies were Sensor and ICT for C4I System and Surveillance and Reconnaissance System and Platform/Structure for Land System, Aeronautical System and Naval Sea System. In contrast, we recognized inadequate technologies, such as Propellant and Material for Missile and Ammunition system and Sensor and ICT for Defense System.  Big data analysis starting from social network analysis has recently emerged as a methodology for analyzing information in unstructured databases and extracting the value of data from them. In particular, existing statistical analysis is optimized for quantitative (static) data analysis, but big data analysis can analyze even non-static data expressed in letters, resulting in many results not only in social science but also in science and technology. In particular, since the amount of information related to science and technology increases explosively as technology development is advanced, big data analysis techniques for analyzing technology trends are an indispensable factor, not an option. Therefore, it has been widely used in the field of science and technology for analyzing and predicting technological trends such as future technology, 4th Industrial Revolution, and nanotechnology.  In the field of social science, we have dealt with the question of centrality as to whether a person is at the center of a human relationship and affects others, the question of connectivity between people and how relationships are connected, and have applied/developed social network analysis as a methodology to confirm it. A network is a structure in which objects are interconnected, and an object is a node that expresses the connection between objects (nodes) as a link. If you switch nodes to "people" and links to "people-to-people relationships," you can analyze the network through social networking methods. The centrality analysis is used as the main methodology for network analysis, which is classified as the aforementioned degree of connectivity, connectivity, and medial centroid, enabling quantitative numerical-based schematics of network components.  The data used in this study was obtained from the Defense Strength Venture Tech-Fin Net in DTiMS.The "Defense Strong and Strong Venture Tech-Fi Net" provides detailed information such as "Company Name", "Company Information", "Holding Technology Name", "Detailed Technology Details", "Usable Areas (Inorganic System Classification/Defense Technology Classification"), "Keywords", and "Registration Date". |

**2-2. reivew**

|  |
| --- |
| I have studied a lot about the importance and utilization of social network analysis through this paper, especially the centrality analysis has recently emerged as a methodology for analysis of unstructured big data, and has been used in various fields ranging from social science areas such as politics, economy, society, culture, etc. to science and technology areas. In science and technology areas, it was also used to investigate/analyze the level of technology in various industrial sectors through the centrality analysis and to establish the technology planning applied thereto.  And although the centrality analysis was not well understood, this opportunity was well organized. Centrality analysis is used as a tool to identify interdependencies between data elements and to quantify the behavior and direction between components by allowing analysis of unstructured data (texts) that were not possible with conventional statistical analysis methods.  Using the above concepts, this study analyzed the relationship between weapon systems and defense technology classification through the centrality analysis of technologies owned by small and medium-sized venture companies in Korea based on the database of Defense Strong Venture Tech-Fi Net operated by DTIMS. I found the results of the actual relationship analysis very interesting.  As a result of the centrality analysis, the "Motorized System" showed the highest centrality of connectivity, and the "Power Support System" showed the highest centrality in proximity and mediated centrality. On the other hand, the "Defense Intelligence System" showed the lowest centrality in all three centrality analyses.  As a result of simultaneous analysis of weapons systems and defense technology classifications, the "Motorized Systems" and "Information and Communication" showed high degree of connectivity centrality, as with individual analysis results. In terms of proximity and mediated centrality, it has been confirmed that "control electronics" technology has a much higher centrality than other technology categories. As shown above, we can see that you are good at analyzing social networks.  I draw based on the centrality analysis results quantified by analyzing unstructured data in the paper above. Through schematics such as 3, 4 and 5, the relationship between elements was visually understood. In particular, while traditional frequency-based analysis methodology simply determines the importance of technology in the order of frequency with the concept of "multiplicative" for the combinations between each classification, the centrality analysis (annuality, proximity, medium) was able to map the macro-relationship beyond the simple frequency by analyzing the relationship data between the analytical elements in the same way as shown in Figure 5.  In addition, the analysis of the social network of the "Defense Strong Venture Tech-Fi Net" confirmed that the technologies possessed by small and medium-sized venture companies in Korea mainly consist of "sensor" and "information and communication" technologies applied to "monitoring/verification systems" and "engineering systems", "fault systems" and "air systems". On the other hand, the development of "promotion" and "materials" technologies applied to "firepower systems" was required, and it was confirmed that the application of "sensor" and "information and communication" technologies was insufficient in "protection systems". As such, I could see that the lack of social network analysis can be accurately analyzed.  I was able to grasp the technological trends in the weapon systems and defense technology classification of technologies possessed by small and medium-sized domestic companies by utilizing the results of this study and the analysis of social networks. Through this, it has been able to analyze strengths and weaknesses of domestic small and medium-sized venture businesses' technologies, and it is expected that they will be used as reference materials for presenting direction when developing technologies for small and medium-sized venture businesses in the future. |

**3. PCA 주성분 분석**

. 논문명 :

. 저자 :

**3-1. summary**

|  |
| --- |
|  |

**3-2. reivew**

|  |
| --- |
|  |

**4. K-means clustering**

. 논문명 :

. 저자 :

**4-1. summary**

|  |
| --- |
|  |

**4-2. reivew**

|  |
| --- |
|  |