हाइपर स्पेक्ट्रल डेटा का निगरानी रहित अधिगम

एल्दो वर्गीस*, सत्यानन्दन टी.वी., विवेकानन्द भारती, ग्रिन्सन जॉर्ज, सोमी कुर्याकोस, मिनी के.जी. और जयशंकर जे.

भा कृ अ नुप- केन्द्रीय समुद्री मात्स्यिकी अनुसंधान संस्थान, कोच्ची, केरल

*संपर्क: eldho.varghese@icar.gov.in

भूमिका

हाइपर वर्णक्रमीय इमेजिंग सेंसर प्रत्येक पिक्सल क्षेत्र के भीतर सामग्री की विकीर्णता को एक बहुत बडी मात्रा में सन्निहित वर्णक्रमीय तरंग दैर्ध्य में मापता है। इसलिए, सतह पर एक दृश्य की सैकड़ों छवियों को उत्पन्न किया जा सकता है। विकीर्णता को हाइपर वर्णक्रमीय डेटा क्यूब डिजिटल रूप में परिवर्तित किया जाता है। एक हाइपर वर्णक्रमीय छवि (क्यूब) में उपलब्ध वर्णक्रमीय जानकारी लक्ष्य वस्तु की प्रकृति को वर्गीकृत करने की सुविधा प्रदान कर सकती है क्योंकि प्रत्येक सामग्री में एक विशिष्ट निश्चित वर्णक्रम होता है और इसका उपयोग सामग्री के एक वर्णक्रमीय हस्ताक्षर के रूप में किया जा सकता है और संभवतः आगे की प्रक्रिया के लिए अतिरिक्त जानकारी एवं पूर्वेक्षण प्रदान करता है। हाइपर वर्णक्रमीय डेटा में अत्यंत समृद्ध वर्णक्रमीय गुण होते हैं, जो वर्गीकरण सटीकता के साथ अधिक विस्तृत श्रेणियों को भेद करने की क्षमता प्रदान करते हैं।

निगरानी रहित अधिगम एक प्रकार की मशीन शिक्षा एल्गोरिथ्म है, जो लेबल की गयी प्रतिक्रियाओं के बिना इनपुट डेटा से युक्त डाटासेट्स के संदर्भ आकर्षित करता है। सबसे आम निगरानी रहित अधिगम की विधि क्लस्टर विश्लेषण है, जो डेटा में छिपे हुए पैटर्न या वर्ग को खोजने के लिए उपयोग किया जाता है।

क्लस्टर विश्लेषण आम तौर पर वर्गों के मामलों को गठबंधन करने की कोशिश में किया जाता है, जब वर्ग सदस्यता विश्लेषण से पहले ज्ञात नहीं होती है। क्लस्टर विश्लेषण व्यक्तिगत वर्गीकरण या अज्ञात वर्गों की वस्तुओं के वर्गीकरण के लिए एक तकनीक है।

किसी भी डेटा का क्लस्टर विश्लेषण करने के लिए चार बुनियादी चरण हैं। ये नीचे दिए गए हैं

1. एक उपयुक्त दूरी माप चुनें
2. एक क्लस्टरिंग एल्गोरिथ्म चुनें
3. समूहों की संख्या निर्धारित करें
4. विश्लेषण को मान्य करें

क्लस्टरिंग तरीके

सामान्यतः क्लस्टरिंग का इस्तेमाल किए जाने वाले तरीके दो सामान्य श्रेणियों में आते हैं।

1. पदानुक्रमिक और
2. गैर पदानुक्रमिक

(क) पदानुक्रमित क्लस्टर विश्लेषण

पदानुक्रमिक क्लस्टरिंग तकनीक या तो एक श्रृंखला के विलय या लगातार प्रभागों की श्रृंखला से आगे बढ़ती है। संचय पदानुक्रमित विधि व्यक्तिगत वस्तुओं से शुरू होती है, इस प्रकार वस्तुओं के रूप में कई समूह हैं। सबसे समान वस्तुओं को पहले वर्गीकृत किया जाता है और ये आरंभिक समूह उनकी समानता के अनुसार विलय कर दिए जाते हैं। फलतः जैसा कि समानता घट जाती है, सभी उपसमूह एक क्लस्टर में जुड़ जाते हैं।

विभाजनकारी पदानुक्रमित तरीके विपरीत दिशा में काम करते हैं। वस्तुओं का एक प्रारंभिक एकल समूह दो उप समूहों में विभाजित है, जैसे कि एक उप समूह की वस्तुएं दूसरों की वस्तुओं से बहुत दूर हैं। इन उपसमूहों को फिर अलग-अलग उपसमूहों में विभाजित किया जाता है। प्रक्रिया तब तक जारी रहती है जब तक वस्तुओं के रूप में कई उपसमूह होते हैं, यानी, जब तक किसी समूह से प्रत्येक वस्तु न हो। दोनों संचय और विभाजनकारी विधि के परिणामों को दो आयामी आरेख के रूप में प्रदर्शित किया जा सकता है जो डेन्डोग्राम नाम से जाना जाता है। यह देखा जा सकता है कि डेन्ड्रोग्राम विलय या विभाजन को दर्शाता है जो कि विभिन्न स्तरों पर किए गए हैं।

अनुबंधन विधियाँ क्लस्टरिंग वस्तु के साथ ही साथ चर वस्तुओं के लिए भी उपयुक्त है। यह सभी पदानुक्रमित संचय प्रक्रिया के लिए सच नहीं है। निम्नलिखित प्रकार के अनुबंधन पर अब चर्चा कर रहे है:

1. एकल अनुबंधन (न्यूनतम दूरी या निकटतम नज़दीक),

2. पूर्ण अनुबंधन (अधिकतम दूरी या दूरतम नज़दीक) और
3. औसत अनुबंधन (औसत दूरी)

इसके अलावा 'वार्ड की विधि', 'सेन्ट्रोइड विधि' जैसे पदानुक्रमिक क्लस्टरिंग तकनीकों के अन्य तरीके भी साहित्य में उपलब्ध हैं ।

(क).(क). पदानुक्रमिक क्लस्टर विश्लेषण में संचय के चरण

N वस्तुओं (वस्तुओं या चर) के समूह के लिए समूहबद्ध पदानुक्रमिक क्लस्टरिंग एल्गोरिथ्म में निम्नलिखित चरण हैं:

1. N क्लस्टर्स से प्रारंभ करें, प्रत्येक में एक ही इकाई और NxN सममित मैट्रिक्स की दूरी (या समानताएं) हो D = {dik}

2. निकटतम (सबसे अधिक समान) क्लस्टरों की जोड़ी के लिए दूरी मैट्रिक्स खोजें। सबसे अधिक समान क्लस्टर U और V के बीच की दूरी duv होनी चाहिए।

3. क्लस्टर U और V विलय करें। नवगठित क्लस्टर (UV) को लेबल करें। दूरी मैट्रिक्स में प्रविष्टियों को अपडेट करें जैसे (क) क्लस्टर U और V के अनुरूप पंक्तियों और कॉलम को हटाने से (ख) क्लस्टर (UV) और शेष क्लस्टर के बीच की दूरी को एक अलग पंक्ति और कॉलम में जोड़ने से।

4. कुल N-1 बार चरण (1) और (3) दोहराएं। (एल्गोरिथम समाप्त होने के बाद सभी वस्तुएं एक क्लस्टर में होंगे)। विलीन होने वाले समूहों की पहचान और स्तर जिस पर विलय हो उसकी दूरी या समानता रिकार्ड करें।

क्लस्टर विश्लेषण के पीछे छिपे हुए मूल विचार अब लिंकेज विधि के एल्गोरिथ्म घटकों को प्रस्तुत करते हुए दिखाए जाते हैं ।

(ख) गैर पदानुक्रमित क्लस्टरिंग विधि

गैर-पदानुक्रमित क्लस्टरिंग तकनीक K समूहों के संग्रह में, चर के बजाय समूह वस्तुओं के लिए बनाया गया है । क्लस्टर की संख्या, K, या तो पहले से निर्दिष्ट किया जा सकता है या क्लस्टरिंग प्रक्रिया के भाग के रूप में निर्धारित किया जा सकता है, क्योंकि दूरी का एक मैट्रिक्स निर्धारित नहीं होता है और बुनियादी डेटा को कंप्यूटर संचालन के दौरान संग्रहीत नहीं करना पड़ता है । पदानुक्रमित तकनीकों की तुलना में बहुत अधिक डेटा समूहों पर गैर पदानुक्रमित तरीकों को लागू किया जा सकता है । गैर पदानुक्रमित तरीकों से या तो (1) समूह में वस्तुओं के प्रारंभिक विभाजन या (2) बीज अंक का एक प्रारंभिक सेट से शुरू होता है जो क्लस्टर के नाभिक बनेगा।

(ख).(क). 'K' का अर्थ क्लस्टरिंग (अफीफी, क्लार्क और मार्ग, 2004)

'K' का मतलब क्लस्टरिंग एक लोकप्रिय गैर पदानुक्रमित क्लस्टरिंग तकनीक है । 'K' के विशिष्ट समूहों के लिए निम्न चरणों में बुनियादी एल्गोरिथ्म प्राप्त होती है:

- डेटा को 'K' प्रारंभिक क्लस्टर में विभाजित करें। इन समूहों की संख्या उपयोगकर्ता द्वारा निर्दिष्ट की जा सकती है या किसी मनमानी प्रक्रिया के अनुसार प्रोग्राम द्वारा चयनित हो सकती है ।

- 'K' समूहों के साधनों या केंद्र की गणना

- किसी दिए गए मामले के लिए, प्रत्येक केंद्र के लिए इसकी दूरी की गणना करें। यदि मामला अपने क्लस्टर के केंद्र के सबसे निकट है, तो उस क्लस्टर में छोड़ दें; अन्यथा, इसे क्लस्टर के लिए पुन: निर्दिष्ट करें जिसका केन्द्रक इसके निकटतम है ।

- प्रत्येक मामले के लिए कदम (3) दोहराएँ ।

- जब तक कोई भी मामलों को पुन: निर्दिष्ट नहीं किया जाता है तब तक चरण (2), (3) और (4) दोहराएं।

डेन्डोग्राम

डेन्डोग्राम को पदानुक्रमित वृक्षारेख या प्लॉट भी कहा जाता है, और निकटता गुणांक के सापेक्ष आकार को दर्शाता है, जिस पर मामलों को जोड़ दिया जाता है । बड़ी दूरी गुणांक या छोटे समानता गुणांक जितना अधिक होगा, उतना ही क्लस्टरिंग में शामिल होना चाहिए, जो संस्थाओं के विपरीत संयोजन करना है, जो कि अवांछनीय हो सकता है । कम दूरी दिखाए जाने वाले मामले बंद हैं, उन्हें एक रेखा के साथ डेन्डोग्राम की बाईं ओर से थोड़ी दूरी पर जोड़कर, एक समानता का संकेत मिलता है । दूसरी तरफ, जब लिंकिंग रेखा डेंड्रोग्राम के दायीं ओर होती है, तो लिंक उच्च दूरी के गुणांक पर होता है, इससे संकेत मिलता है कि मामले / समूहों को समुच्चयित किया गया था, हालांकि बहुत कम समान है ।

दूरी उपाय कार्यवाही

आमतौर पर आकस्मिक वर्णक्रमीय समानता / असमानता का आकलन करने के लिए उपयोग किए जाने वाले कुछ दूरी उपाय निम्नानुसार हैं:

1. स्पेक्ट्रल समानता सूचकांक या स्पेक्ट्रल सहसंबंध या स्पेक्ट्रल कोण मैपर (एस ए एम)
2. स्पेक्ट्रल व्यतिरेक कोण
3. स्पेक्ट्रल सूचना अंतर (एस आई डी)
4. स्पेक्ट्रल अवशोषण सूचकांक (एस ए आई)
5. यूक्लिडियन दूरी
6. महालनोबिस D2
7. सिटी ब्लॉक दूरी

सत्यापन

- सापेक्ष क्लस्टरिंग सत्यापन, जो समान एल्गोरिथम के लिए अलग-अलग पैरामीटर मानों को बदलकर क्लस्टरिंग संरचना का मूल्यांकन करता है । (उदाहरण के लिए: K समूहों की संख्या को अलग करना) । यह आम तौर पर समूहों की इष्टतम संख्या का निर्धारण करने के लिए उपयोग किया जाता है ।

- बाह्य क्लस्टरिंग सत्यापन, जिसमें एक क्लस्टर विश्लेषण के परिणाम बाह्य रूप से ज्ञात परिणाम, जैसे बाह्य रूप से प्रदान किए गए कक्षा लेबल, की तुलना में शामिल होते हैं । चूंकि हम "सही" क्लस्टर नंबर को पूर्व में जानते हैं, इसलिए यह दृष्टिकोण मुख्य रूप से किसी विशिष्ट डाटासेट के लिए सही क्लस्टरिंग एल्गोरिथ्म चुनने के लिए उपयोग किया जाता है ।

- आंतरिक क्लस्टरिंग सत्यापन, जो बाहरी सूचना के संदर्भ के बिना क्लस्टरिंग संरचना की भलाई का मूल्यांकन करने के लिए क्लस्टरिंग प्रक्रिया की आंतरिक जानकारी का उपयोग करती है । इसका उपयोग किसी बाह्य डेटा के बिना समूहों की संख्या और उचित क्लस्टरिंग एल्गोरिथम के आकलन के लिए भी किया जा सकता है ।

- क्लस्टरिंग स्थिरता सत्यापन, जो आंतरिक सत्यापन का एक विशेष संस्करण है । यह एक क्लस्टरिंग परिणाम की निरंतरता का मूल्यांकन जिसके द्वारा प्रत्येक स्तंभ हटाए जाने के बाद प्राप्त समूहों के साथ तुलना करके करता है ।

सारांश

वास्तविक समूह सदस्यता का ज्ञान अज्ञात है, तो अन्वेषित शिक्षा समूह के मामलों को वर्गीकृत करने के लिए एक अनुमानन तकनीक है । जब तक अंतर्निहित समूह के बीच काफी अंतर नहीं होता है, निगरानी रहित अभिगम के साथ बहुत स्पष्ट परिणाम की अपेक्षा करना यथार्थवादी नहीं है । विशेष रूप से अगर अवलोकनों को गैर-रेखीय तरीके से वितरित किया जाता है, तो अलग-अलग समूहों को प्राप्त करना मुश्किल हो सकता है । क्लस्टर विश्लेषण बाहरी कारकों के प्रति काफी संवेदनशील है । क्लस्टर कार्यक्रम चलाने से पहले डेटा को ध्यान से जांचना चाहिए।

संसदीय राजभाषा समिति द्वारा निरीक्षण

संसदीय राजभाषा समिति की दूसरी उप समिति ने दिनांक 22.01.2018 को सी एम एफ आर आइ मुख्यालय, कोचीन की राजभाषा गतिविधियों का निरीक्षण किया। निरीक्षण समिति में डॉ. प्रसन्न कुमार पाटसाणी, सांसद (लोक सभा), डॉ. सुनील बलिराम गायकवाड़, सांसद (लोक सभा), डॉ. लक्ष्मी नारायण यादव, सांसद (लोक सभा), डॉ. सत्येन्द्र सिंह, वरिष्ठ अनुसंधान अधिकारी, डॉ. विकास वर्मा, हिन्दी अधिकारी, श्रीमती नीरजा, अनुसंधान सहायक और श्री अब्दुल मोहीब, सहायक सम्मिलित थे।

समिति ने संस्थान की राजभाषा कार्यविधियों का निरीक्षण करके संतोष प्रकट किया। निरीक्षण बैठक के दौरान संस्थान के प्रकाशनों / राजभाषा पुरस्कारों, शील्डों तथा ट्रोफियों का प्रदर्शन भी आयोजित किया गया।